主査：池田哲夫　　教授
副査：渡邉貴之　准教授
副査：風間一洋　　教授
指導教員：斉藤和巳　　教授

統計的機械学習に基づく
多種データ統合利活用技術に関する研究

学籍番号 1424505

山岸祐己

平成29年 1月 27日

論文要旨

本論文では，多様な環境において生成された大規模データの特徴や問題点を露呈させ，それら大規模データの利活用の一助となるような汎用技術を開発する．

詳細には，まず，基礎的なデータ処理を目的として，様々な形式の時系列データに対する新たな異常検出手法を提案する．提案異常検出手法は，短期的な異常（バースト）だけではなく，長期的な異常（潜在的変化）も混在していることを仮定した区間分割手法である．この手法の応用として，データの時期的な信頼性を考慮したデータ評価が可能かどうかを検証した．

続いて，提案した区間分割法によるデータ評価法をもとに，時期的な信頼性に対する別のアプローチである時間減衰関数と，データの位置情報を利用した空間減衰関数を導入したデータカテゴリの評価モデルを構築する．さらに，評価方法を巨視的な見地まで発展させるべく，データオブジェクトをノードとした完全ネットワークモデルを提案し，複数のノード指標からデータカテゴリの特徴を見る．最後に，提案ネットワークモデルの応用として，データカテゴリをノードとした時系列データを分析した．

これら問題に対する代表的な既存手法，およびナイーブな手法との比較実験結果から，複数属性を有する時系列データ，位置情報と時刻情報を有する時空間データ，バッチ処理で取得されたデータに対しての，提案手法の有効性と有用性を確認した．

Abstract

In this thesis, for the purpose of utilising large-scale data, we develop general purpose technologies that expose features and problems of such large-scale data generated in various environments.

In detail, we first propose a new anomaly detection method for various types of time series data for the purpose of basic data processing. The proposed anomaly detection method is an interval segmentation method assuming that not only short-term anomalies (bursts) but also long-term anomalies (latent changes) are mixed. As an application of this method, we evaluated time-series data in consideration of temporal reliability using the segmentation result of the method.

Subsequently, based on the time-series data evaluation method by the proposed interval segmentation method, we construct an evaluation model of data categories introducing a temporal decay function which is another approach to temporal reliability and a spatial decay function related to the position information. Furthermore, in order to expand the evaluation method as macroscopic analysis, we propose a complete network model in which the data objects are nodes and clarify the features of the data categories from the various node evaluations. Finally, as an application of the proposed network model, we also dealt with time series analysis when the nodes are the data categories.

From the results of comparative experiments using representative conventional methods and naive methods for these problems, we confirmed the effectiveness and usefulness of the proposed method for time-series data with multiple attributes, spatiotemporal data with position and timestamps, and data acquired by batch processes.

凡例

数式，図表などは情報処理学会の規定に準拠する．

目次

第1章

序論

本章では，本論文の研究背景および研究目的について言及する．

1.1 研究背景

近年，現実世界とインターネットの両環境において爆発的なデータの増加が起こっており，それら大規模データが世に溢れている現状と計算機の能力向上に伴って，統計的機械学習の技術研究は更なる発展を遂げている．一般に，人々が大規模データに触れる機会は多くなっているが，大規模データに対するリテラシーがそれに伴って世の中に普及しているとは限らない．現に，人々に提供するナイーブなソーシャル情報の項目数を増やすと，個々の意思決定に多大な影響を与え，市場の不平等性を大いに増加させることが，大勢の参加者による実験 [1] によって既に分かっている．ここで，ナイーブなソーシャル情報とは，例えば，訪問者数，購入者数，視聴者数，ダウンロード数，コメント又はレビュー数といった，大規模データにおける単純な情報をまとめたものである．もし，このような現状が今後も続くのであれば，大規模データの中でも，既にナイーブなソーシャル情報において目立っているデータは，常に利用される可能性が高くなり，相対的にその他のデータや新規で追加されたデータは利用される可能性が低くなってしまう．

現在の実用上，上記のことは大きな問題では無いかもしれないが，特に推薦システムの分野では，これらが潜在的な問題として扱われている [2]．この潜在的な問題というのは，セレンディピティ (serendipity) という言葉に関するもので [3]，この言葉の解釈としては，予想外のものを発見すること，探しているものとは別の価値があるものを偶然見つけること，ふとした偶然をきっかけに幸運をつかみ取ること，などがある．つまり，大規模データ内の有益なデータや情報は，ナイーブなソーシャル情報を頼りにしていると見過ごしてしまう恐れがあり，それらに辿り着けるかどうかは個々人の能力に依存してしまうという問題である．コンピュータ科学界隈では，セレンディピティの概念定義や，セレンディピティを向上させる方法について，様々な研究がなされている [4, 5, 6, 7, 8]．

したがって，大規模データにおける情報の優劣における不平等性を抑制し，有益な情報を得るための

一助となるような汎用技術の開発は重要であると言える．そのためにはまず，データの特徴や問題点を知るための基礎的な処理として，汎用的に使える異常検出手法が必要となる．時系列データにおける異常検出手法は，Kleinberg のバースト検知 [9] をはじめとして幅広く研究がなされており [10, 11]，近年では異常（バースト）の数理モデル化の研究も進んでいる [12]．更に，時系列データに対する別のアプローチとしては，新しいアイテムと古いアイテムを平等に評価することを目的として，時間減衰関数 [13, 14] が頻繁に用いられている．実際，推薦システムにおいても，時間減衰関数を用いたモデルが提案されている [15]．そして，巨視的なデータ評価をするための処理としては，大規模データの複雑ネットワーク化が挙げられる．元来，大規模なデータを利用した複雑ネットワークの構造や機能に関する研究は，社会学，生物学，物理学，コンピュータ科学等の様々な分野で注目されている [16]．中でも，様々な側面から重要ノード群を発見することは，ネットワーク分析において基礎的な問題とされており [17]，Web 情報検索の分野では，PageRank [18] と HITS [19] によるノードランキングが広く用いられている．

1.2　研究目的

本論文の目的は，上記の研究背景における種々の問題を扱うべく，大規模データの利活用の一助となるような汎用技術を開発することである．ここでの汎用技術とは，多様な環境において生成された大規模データに対応することができ，使用する際には，事前のデータ処理や詳細な設定がなるべく不要なものを指す．例えば，時系列の異常検出においては，データの観測時刻が連続であっても離散であっても扱うことができ，事前に設定するパラメータもせいぜい検出感度だけで済むようなものが望ましい．また，大規模データの巨視的分析として幅広い利活用が期待できるデータオブジェクト評価，又はそれを発展させたデータカテゴリ評価においては，複雑ネットワークの先行研究技術を用いたモデルが，有用性の高い評価モデルとして期待できる．特に，位置情報を有している大規模データは，現実空間をネットワーク化することにより，より現実に即したデータ評価が可能となる．

1.3　論文構成

本論文は「統計的機械学習に基づく多種データ統合利活用技術に関する研究」と題し，多様な環境において生成された多種データに対する汎用技術についての一連の研究をまとめたものであり，本論文は 8章からなる．

第 1章の序論に続いて，第 2章では，多種データに対して汎用的に利用できる異常検出手法として提案した区間分割法について述べる．区間分割法は，属性が複数含まれていたり，観測時刻間隔が一定だったりするような多種データを扱うことを前提として，多様な区間長の異常を検出できるよう，データ属性の確率分布に対する尤度最大化と尤度比検定に基づいて構築したものである．ここでは，代表的

な異常検出手法との比較実験において，区間分割法のパラメータ設定の容易さと，結果の明快さについて示す．

第 3章では，第 2章で提案した区間分割法の大規模データへの応用について述べる．ここでは，大規模なレビューサイトのデータを用いて，提案した区間分割法がレビューサイト特有の問題をどこまで扱えるかを検証する．更に，高次元属性を持つようなデータへの応用についてもここで述べる．

第 4章では，第 2章で提案した区間分割法の結果の利用方法について述べる．分割した区間において，特に区間の異常度が高いものは異常区間として評価の対象外としたり，直近の分割区間だけを信頼区間として評価対象としたりすることがここにおける目的である．ここでは，区間を z-score として扱う方法と，それを利用したランキング手法について提案する．

第 5章では，位置情報と時刻情報が正確であるデータを扱うことを前提として，第 4章で提案した z-score を時空間モデルとして拡張する．区間分割法を利用したアイテム評価は，各アイテム依存の問題を露呈させることには向いているが，各アイテムの z-score もそれら個々の問題に応じて急激に変化してしまうため，ここでは一度区間分割をしない状態を考える．ただし，区間分割法のときと同様，データの時期的な信頼性を考慮して，z-score に時間的信頼減衰関数を導入する．さらに，位置情報を利用した空間的信頼減衰関数も導入し，順位統計量に基づいたカテゴリ評価の観点から z-score の平等性を向上させることを試みる．

第 6章では，第 5章で述べた時空間モデルをさらに発展させたネットワークモデルについて述べる．このネットワークモデルは，各データオブジェクトを完全ネットワークのノードとして扱い，時空間情報をネットワーク内の移動情報として利用するものである．ただし，その移動情報はノード間の距離とノードの人気度（知名度・認知度）に依存していると仮定し，距離に関するパラメータと人気度に関するパラメータは機械学習的手法で推定する．ここでは，このネットワークモデルにおけるノードの評価指標と，それらを使ったカテゴリ評価方法について提案する．

第 7章では，第 6章のネットワークモデルの応用について述べる．例えば，位置情報が無かったり，時刻情報もバッチ処理的に統一されているようなデータにおいては，ネットワーク内の移動という概念は使えない．しかし，データオブジェクトをある種のカテゴリに分けたときのカテゴリ間類似度は，大抵のデータにおいて計算することができるため，それを使った巨視的かつ汎用的なネットワークモデルを提案する．応用ネットワークモデルでは，カテゴリがネットワークのノードとなるため，第 6章で提案したノードの評価指標そのものがカテゴリ評価指標として用いることができる．ここでは，それらカテゴリ評価指標の時系列的変化に着目し，クラスタリングによってカテゴリに関する規則性や重要性を見出すことを試みる．

第 8章は結論であり，研究成果をまとめている．

第2章

区間分割法：関連研究との比較

この章では，多種データに対して汎用的に利用できる新たな異常検出手法について述べる．特に，未知の環境において生成されたデータに対しては，事前の異常検出によって，その環境における特性や問題点を露呈させることが重要となる．

なお，区間分割法を扱う第 2章から第 4章までは，数式や解法において，共通の定義を使用している．

2.1 はじめに

時系列データにおける異常検出手法は，Kleinberg のバースト検知 [9] をはじめとして幅広く研究がなされており [10, 11]，近年では異常（バースト）の数理モデル化の研究も進んでいる [12]．Kleinberg のバースト検知は，パラメータを多様に調整できることを考慮すると汎用性が高いように思えるが，バースト状態を検知する感度のパラメータ(γ)を決定した時点で想定されるバースト（異常）区間の長さもある程度固定されてしまうため，想定よりも長期的な異常，すなわち潜在的な異常を検出できない恐れがある．更に，これらのバースト検知や異常検出手法 [9, 10, 11]は，一般的に 1-属性の情報を扱うために設計されているので，複数属性の情報をもつ時系列データに対しては属性毎に独立して適応する必要があり，複数属性間の関係を考察することが難しい．例えば，ある2-属性がほぼ同時期にバーストを示したとしても，そのバーストの度合いは各属性によって相対的に決められているため，属性間を相対的に比較することができない．また，これらの手法は観測データの時刻間隔に依存するところが大きいため，観測時刻間隔がもともと一定になっているデータや，観測時刻間隔があまり変化しないデータには向いていない．

よって本章では，上記の異常検出手法が扱うようなデータ形式に加え，属性が複数含まれていたり，観測時刻間隔が一定だったりするような多種データを扱うことを前提として，多様な区間長の異常（バースト）を検出する手法を提案する．提案手法は，代表的なバースト検知の研究 [9, 20] と同様の回顧的 (Retrospective) な観点から，データ属性の確率分布に対する尤度最大化と尤度比検定に基づいて考案したものである．Kleinberg の手法におけるパラメータは，バーストのスケーリングとバースト検知の

感度の2つを要するのに対し，提案手法におけるパラメータは，異常検出の感度を決定するものただ1つであるため，汎用性が高い設計となっている．更に，感度のパラメータを固定した状態で，提案手法の方が様々な区間長の異常に対応できることを人工データによる比較実験で示す．

2.2　問題設定

J-属性の情報を持つ時系列データについて，それらの確率分布における変化の異常と，異常の持続区間を推定することを考える．n 番目の観測ステップ t_n での属性情報を s_n とすると，観測された時系列データ $\mathcal{D}$ は

$$\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}, \tag{2.1}$$

のように表せる．ここで，$s_n \in \{1, \cdots, J\}$ である．便宜上，s_n は属性 $j \in \{1, \cdots, J\}$ の J-次元ベクトルダミー変数として

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

のように変換する．属性 j が出現する確率 p_j が多項分布に従っていると仮定すると，時系列データ $\mathcal{D}$ に対する対数尤度は，出現確率のパラメータベクトル $\boldsymbol{p} = \{p_1, \cdots, p_J\}$ によって

$$\mathcal{L}(\mathcal{D}; \boldsymbol{p}) = \sum_{n=1}^{N} \sum_{j=1}^{J} s_{n,j} \log p_j, \tag{2.3}$$

のように計算できる．式(2.3)の最尤推定量は

$$\hat{p_j} = \frac{\sum_{n=1}^{N} s_{n,j}}{N}, \tag{2.4}$$

のように与えられる．

ここからは，このパラメータベクトル $\boldsymbol{p}$ が，確率分布の変化に従って K 箇所で分割された階段関数の形をとることを考える．すなわち，$k \in \{1, \cdots, K\}$ 番目の分割点の時刻 T_k $(t_1 < T_k < t_N)$ においてパラメータベクトルが $\boldsymbol{p}_k$ から $\boldsymbol{p}_{k+1}$ に切り替わることを仮定する．K 個の分割点を持つ集合 $\mathcal{C}_K = \{T_1, \cdots, T_K\}$ とし，便宜上 $T_{k-1} < T_k, T_0 = t_1, T_{K+1} = t_N$ とする．更に，$\mathcal{C}_K$ による $\mathcal{D}$ の分割を

$$\mathcal{D}_k = \{n; T_{k-1} < t_n \leq T_k\}, \tag{2.5}$$

すなわち

$$\mathcal{N} = \{1, 2, \cdots, N\} = \{1\} \cup \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{K+1}, \tag{2.6}$$

とし，$|\mathcal{D}_k|$ は $(T_{k-1}, T_k]$ における観測ステップ数とする．ここで，任意の $k \in= \{1, \cdots, K+1\}$ について $|\mathcal{D}_k| \neq 0$ ，つまり少なくとも一つの $t_n \in \mathcal{D}_k$ が存在することを条件とする．この時点で，この分

割点検出問題は，部分集合 $\mathcal{C}_K \subset \mathcal{T}$ の探索問題ということになる．ここで，$\mathcal{T}$ は観測ステップの集合 $\mathcal{T} = \{t_1, t_2, \cdots, t_N\}$ である．

分割点 $\mathcal{C}_K$ によって与えられる $\mathcal{D}$ の対数尤度は，パラメータベクトル集合 $\boldsymbol{P}_{K+1} = \{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{K+1}\}$ を用いて

$$\mathcal{L}(\mathcal{D}; \boldsymbol{P}_{K+1}, \mathcal{C}_K) = \sum_{k=1}^{K+1} \mathcal{L}(\mathcal{D}_k; \boldsymbol{p}_k), \tag{2.7}$$

のように計算できる．つまり，式(2.7)の k と j についての最尤推定量は

$$\hat{p}_{k,j} = \frac{\sum_{n \in \mathcal{D}_k} s_{n,j}}{|\mathcal{D}_k|}, \tag{2.8}$$

となる．これらを式(2.7)に代入すると，

$$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, \mathcal{C}_K) = \sum_{k=1}^{K+1} \sum_{n \in \mathcal{D}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}, \tag{2.9}$$

が導ける．従って，この分割点検出問題は，式(2.9)を最大化する $\mathcal{C}_K$ の探索問題に帰着できる．しかし，式(2.9)では分割点集合 $\mathcal{C}_K$ の導入によってどれだけ対数尤度が改善し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，分割を考慮しない，すなわちパラメータベクトルの変化が無いことを仮定したときの対数尤度からの改善度合いを評価することが重要となるため，尤度比最大化問題として目的関数を構築する．もし，パラメータベクトルに一切の変化がない，すなわち $\mathcal{C}_0 = \emptyset$ とするならば，式(2.9) は

$$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, \mathcal{C}_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{1,j}, \tag{2.10}$$

となる．ここで，

$$\hat{p}_{1,j} = \frac{\sum_{n \in \mathcal{N}} s_{n,j}}{|\mathcal{N}|}, \tag{2.11}$$

である．よって，K 個の分割点を持つ場合と分割点を一切持たない場合との対数尤度比は

$$\mathcal{LR}(\mathcal{C}_K) = \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, \mathcal{C}_K) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, \mathcal{C}_0), \tag{2.12}$$

のように与えられる．最終的に，この分割点検出問題は $\mathcal{LR}(\mathcal{C}_K)$ を最大化する $\mathcal{C}_K$ の探索問題に帰着できる．

2.3　分割点検出法

式(2.12)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，ここでは任意の K について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，更にそれらを組み合わせた提案解法について説明する．

2.3.1　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分割の繰り返しである．つまり，既に選択された $(k-1)$ 個の分割点 $\mathcal{C}_{k-1}$ を固定したまま k 番目の分割点 T_k を $\mathcal{C}_{k-1}$ に新たに追加することを繰り返す．一般的に，$2(\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1}))$ 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，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.　$k=1, \mathcal{C}_0=\emptyset$ のように初期化する．

A1-2.　$T_k = \mathrm{argmax}_{t_n \in \mathcal{T}}\{\mathcal{LR}(\mathcal{C}_{k-1} \cup \{t_n\})\}$ を探索する．

A1-3.　$\mathcal{C}_k = \mathcal{C}_{k-1} \cup \{T_k\}$ のように更新する．

A1-4.　もし $2(\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1}))$ 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，$\mathcal{C}_K$ 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 $\mathcal{C}_k$ の各分割点は，$T_{i-1} < T_i$ ($i=2,\cdots,k$) 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 にたい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，先程も説明したように，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．

2.3.2　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 $\mathcal{C}_K$ から始まり，分割点の改善を1つずつ試みるものである．つまり，分割点 T_k を一度取り去り，残った $\mathcal{C}_K \setminus \{T_k\}$ を固定して，よりよい尤度を得られる 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，全ての k $(k=1,\cdots,K)$ に対して分割点の置換が行われない，すなわち，全ての 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.　$T'_k = \mathrm{argmax}_{t_n \in \mathcal{T}}\{\mathcal{LR}(\mathcal{C}_K \setminus \{T_k\} \cup \{t_n\})\}$ 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 $\mathcal{C}_K = \mathcal{C}_K \setminus \{T_k\} \cup \{T'_k\}$ の

ように更新する．

A2-4.　もし $h = K$ ならば $\mathcal{C}_K$ 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が経験的に分かっている [21].

2.3.3　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，逐次的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 $\mathcal{C}_K$ を得る．

C2.　A2 で $\mathcal{C}_K$ を改善する．

となる．以降この解法を 従来解法 *(Conventional)* と呼ぶ．確かに，これだけでも十分な近似解が期待できるが，分割点数 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分割点数のまま局所改善を行ってしまう恐れが大いにある．したがって，不必要な分割点は極力追加せず，且つ必要な分割点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み合わせを提案する．提案解法 *(Proposed)* 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，$\mathcal{C}_k$ を $\mathcal{C}_K$ として出力する．

P3.　$\mathcal{C}_K$ を A2 で改善し，改善した $\mathcal{C}_K$ を $\mathcal{C}_k$ として出力する．

P4.　A1-5 から処理を再開させ，ステップ P2 へ戻る．

この手順では，分割点が追加される度に局所探索法アルゴリズムを行うため，更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が経験的に分かっている [21]．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，これが分割点数を大きく左右するため，本手法における感度のパラメータということになる．

2.4　実験

2.4.1　人工データによる比較実験

ここでは，隠れマルコフモデルを用いた Kleinberg([9]) の手法を技術水準とし，提案手法との比較実験を行う．実験で用いるのは 3-属性をランダムに発生させた人工時系列データで，出現確率を階段関数の形で変化させたものである．真の分割点数は $K = 3$ であり，異常（バースト）区間は $\mathcal{D}_2$ と $\mathcal{D}_3$ とした．異常区間の両端部分の観測ステップ数を $|\mathcal{D}_1| = 10000$, $|\mathcal{D}_4| = 10000$ で固定したまま，異常区間の観測ステップ数は5パターン(100, 500, 1000, 5000, 10000)に変化させ，それぞれ100サンプルずつ生成した．Kleinberg の手法のスケーリングパラメータ（今回は $s = 1.5$ ）に則って，出現確率の変化は表 2.1のように設定した．なお，Kleinberg の手法の感度のパラメータは $\gamma = 1.0$, 提案手法のアルゴリズムにおける χ^2 検定の危険率（感度のパラメータ）は $p = 0.0001$ で固定した．

表2.1　3-属性の出現確率設定

	$p_{*,1}$	$p_{*,2}$	$p_{*,3}$
$\boldsymbol{p}_1$ $(\mathcal{D}_1)$	1/3	1/3	1/3
$\boldsymbol{p}_2$ $(\mathcal{D}_2)$	2/4	1/4	1/4
$\boldsymbol{p}_3$ $(\mathcal{D}_3)$	6/8	1/8	1/8
$\boldsymbol{p}_4$ $(\mathcal{D}_4)$	1/3	1/3	1/3

各手法による異常検出結果を基にした推定出現確率と，真の出現確率との絶対誤差の比較を図 2.1 から 2.4に示す．図より，提案手法は極端に短い異常区間 $|\mathcal{D}_2| = |\mathcal{D}_3| = 100$ においては上手く機能していないが，それ以外の場合においての性能は良好である．逆に，Kleinberg の手法は，異常区間が長くなってくると上手く機能しないという特徴が見られる．感度のパラメータは固定しているため，パラメータを変化させてさらに検証する余地はあるが，異常区間の長さが予測できない状況を考慮すると，提案手法のほうが多様な区間長の異常に適応できる可能性が高いと言える．更に，各手法に要した計算時間の比較 (Intel(R) Xeon(R) X5690 @3.47GHz) を図 2.5,2.6に示す．図より，提案手法は個々のサンプルの問題に依存して計算時間の変動が大きいが，全体的な計算時間については，全ての場合において提案手法の方が短いことがわかる．

参考までに，各異常区間における両手法の代表的な結果を図 2.7 から 2.16 に示す

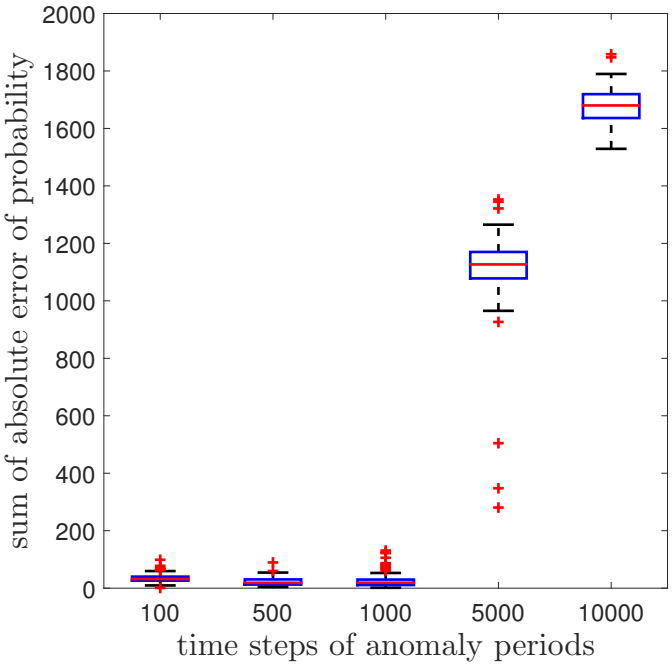

図2.1　真の出現確率との絶対誤差の比較 (Kleinberg の手法)

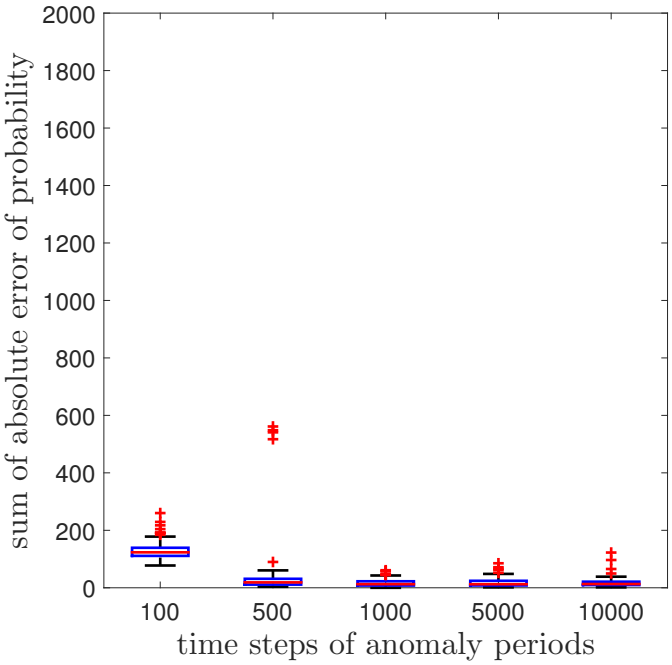

図2.2　真の出現確率との絶対誤差の比較 (提案手法)

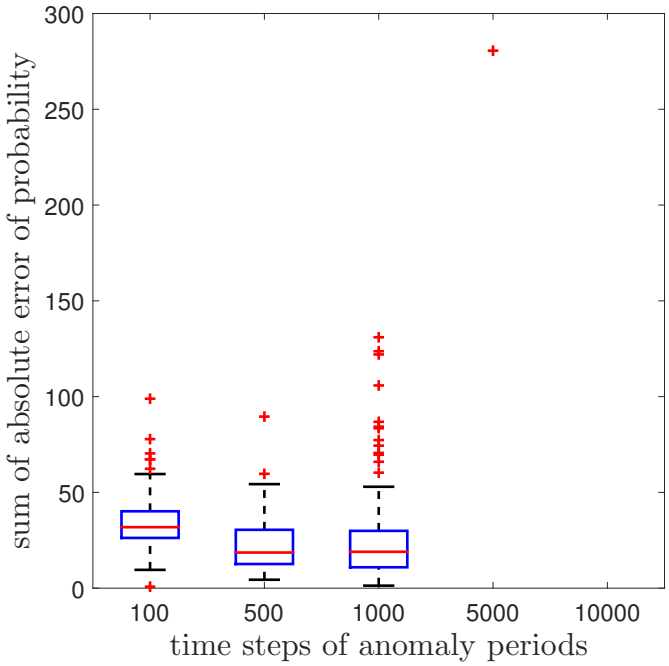

図2.3　真の出現確率との絶対誤差の比較の詳細 (Kleinberg の手法)

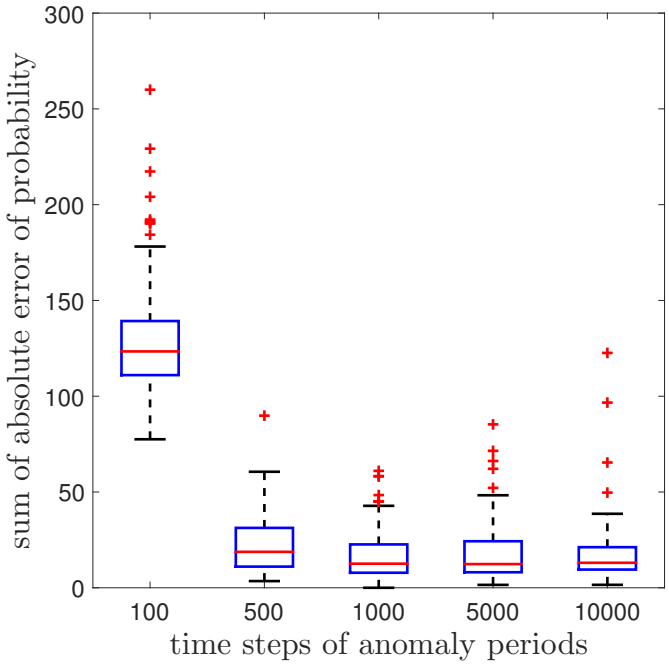

図2.4　真の出現確率との絶対誤差の比較の詳細 (提案手法)

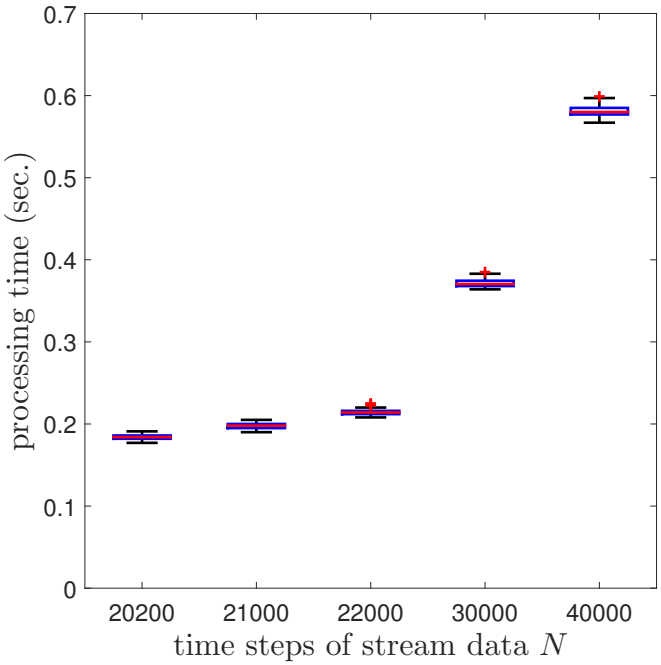

図2.5　計算時間の比較 (Kleinberg の手法)

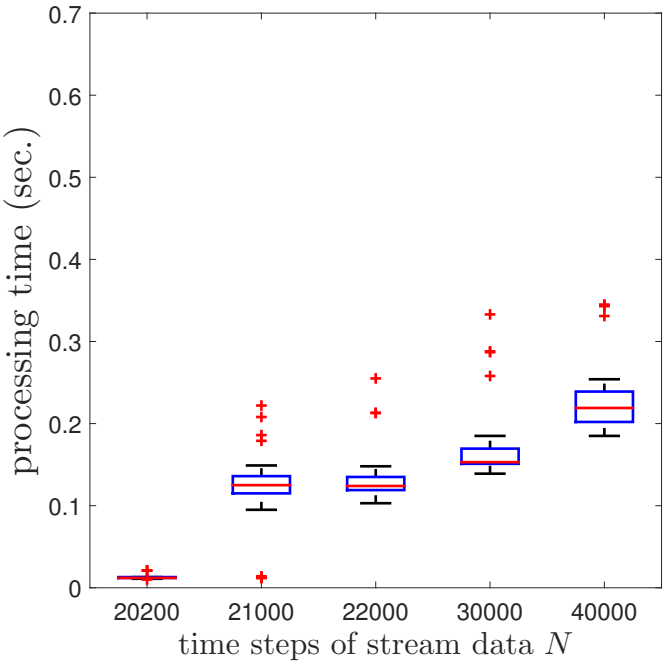

図2.6　計算時間の比較 (提案手法)

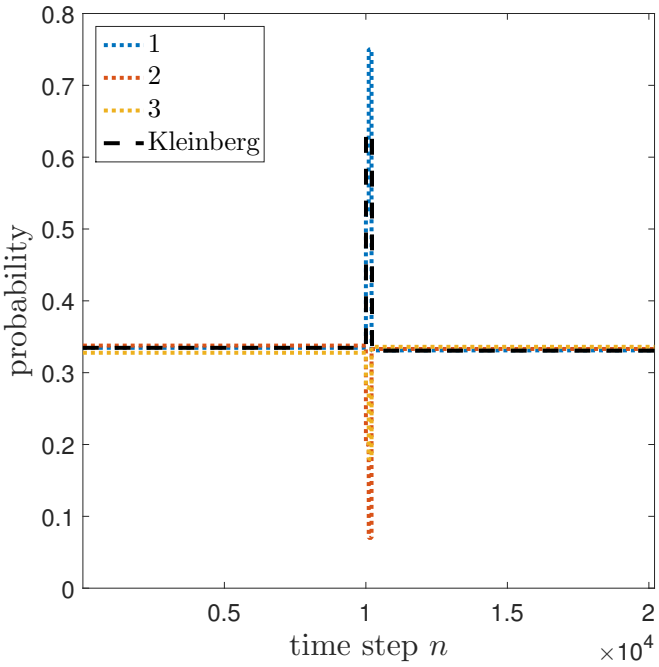

図2.7　異常区間の観測ステップ = 100 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

2.4.2　現実データによる比較実験

現実データとして，国内の大規模化粧品レビューサイト“@cosme” [*1] から，各アイテムにつけられた今までのレビューの点数情報を取得し，最も多くのレビューを有する2アイテム (“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” と “Conditioner Essential (Albion)”) に対して Kleinberg の手法と提案手法を適応した．図 2.17，2.18 に2アイテムのレビュー点数の移動平均（計算範囲100観測ステップ）を示す．図から分かるように，観測ステップ数は人工データのときとほぼ同様となっており，“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” は観測時期による平均点数の変動が大きく，“Conditioner Essential (Albion)” は観測時期による平均点数の変動が小さい．人工データのときと同様，Kleinberg の手法のスケーリングパラメータとバースト感度パラメータは $s = 1.5$ と $\gamma = 1.0$, 提案手法の異常検出の感度パラメータ (χ^2 の危険率) は $p = 0.0001$ とした．なお，@cosme のレビュー点数の範囲は0 から 7 ($j = 1, \cdots, 8$ として考える) であるため，提案手法における自由度は $J - 1 = 7$ となる．まず，図 2.19 から 2.22 に Kleinberg の手法を適応したときの結果を示す．図 2.19，2.21 は各レビュー点数のバーストレベルをそのままプロットしたものであり，

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

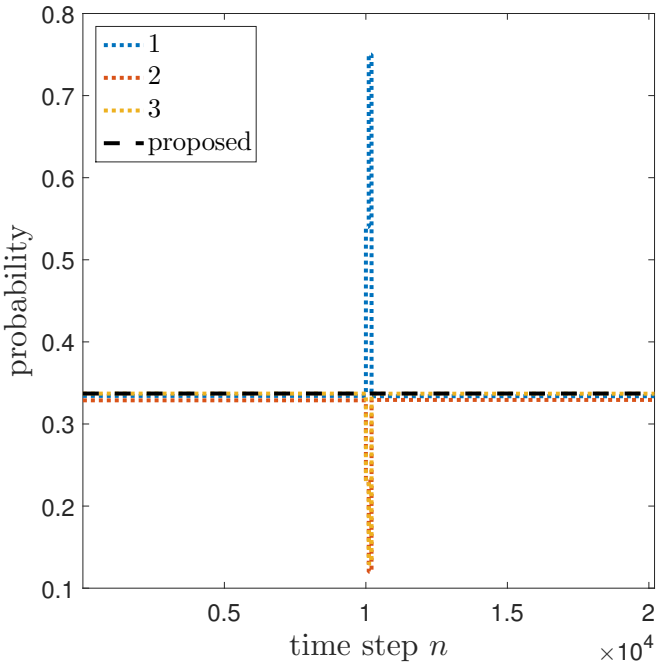

図2.8　異常区間の観測ステップ = 100 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

図 2.20，2.22 はそれらバーストレベルを確率推移として表したものである．図より，Kleinberg の手法では，各レビュー点数の密度がいつ高くなったかということは分かりやすいが，それらの情報がレビュー点数毎で独立しているため，全体的な傾向の流れを把握することは難しいように思える．さらに，“Conditioner Essential (Albion)” の方はレビュー点数の出現傾向の変動が少ないため，序盤の僅かな部分しかバーストとして認識されず，他の部分はノイズの扱いとなっていることが分かる．次に，図 2.23 と　2.25 に提案手法を適応したときの結果を示す．図 2.23より，“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” のデータにおいて，提案手法は適切な分割数と様々な分割区間長によって，レビュー得点の確率分布における変化を表現できているように見える．このときの分割結果に基づいた平均点数の推移は図 2.24となり，平均点数の階段関数への変換としても上手く機能していることが分かる．また，レビュー点数の出現傾向の変動が少ない “Conditioner Essential (Albion)” のデータにおいても，図 2.25 のように最低限の出現傾向の変化は認識できており，このときの分割結果に基づいた平均点数の推移（図 2.26）も自然な出力となっている．更に，提案手法の結果については，分割点が集中している期間や短い区間は得点分布の信頼性が有意に低く，長い区間は得点分布の信頼性が有意に高いということも自然に考えられる．

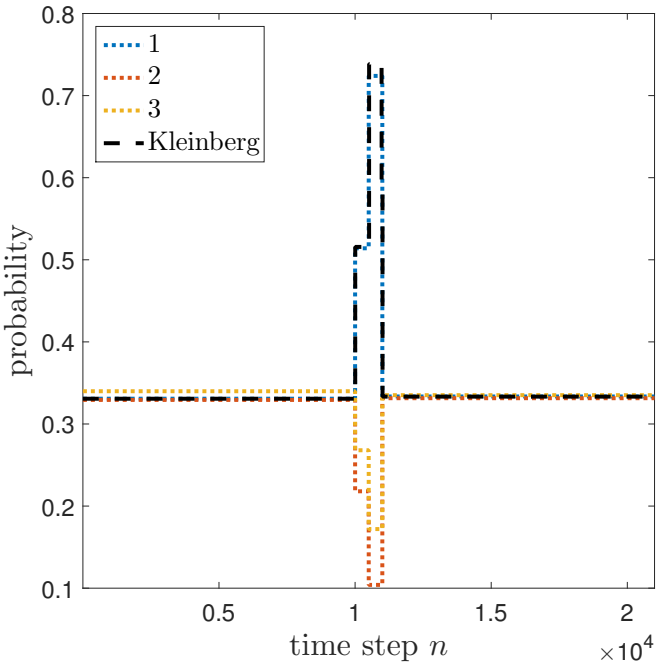

図2.9　異常区間の観測ステップ = 500 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

2.5　おわりに

代表的な異常（バースト）検出手法特有の問題を補うため，属性が複数含まれていたり，観測時刻間隔が一定だったりするような多種データを扱うことを前提として，多様な区間長の異常（バースト）を検出する手法を提案した．本章では，データ属性の確率分布に対する尤度最大化と尤度比検定に基づいた問題設定を行い，その問題を高速且つ高精度に解くための解法を提案した．人工データを使った Kleinberg の手法との比較実験では，両手法の異常検出に関する感度のパラメータを固定した状態で，提案手法の方が様々な区間長の異常に対応できることを示した．今回の比較実験では，計算時間においても提案手法が僅かに優れている結果となった．現実データを使った実験では，レビュー点数の分布変化の視覚化に成功し，得点分布の信頼性指標における有用性を示した．

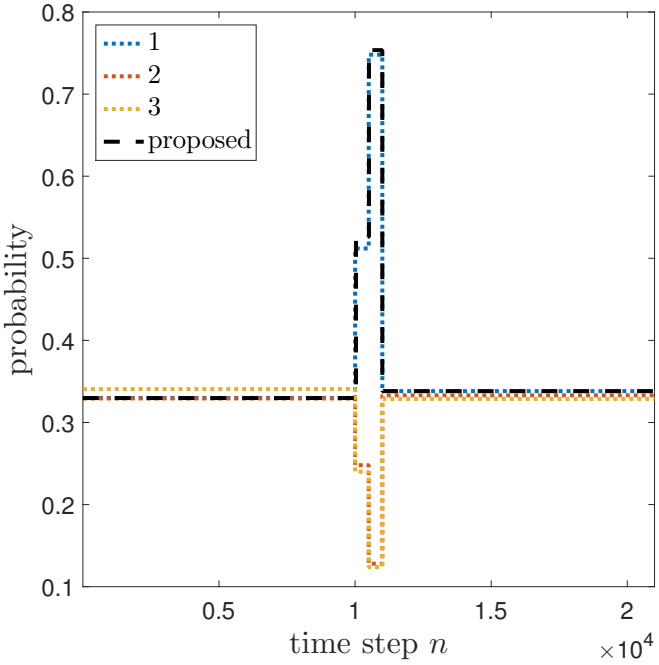

図2.10　異常区間の観測ステップ = 500 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

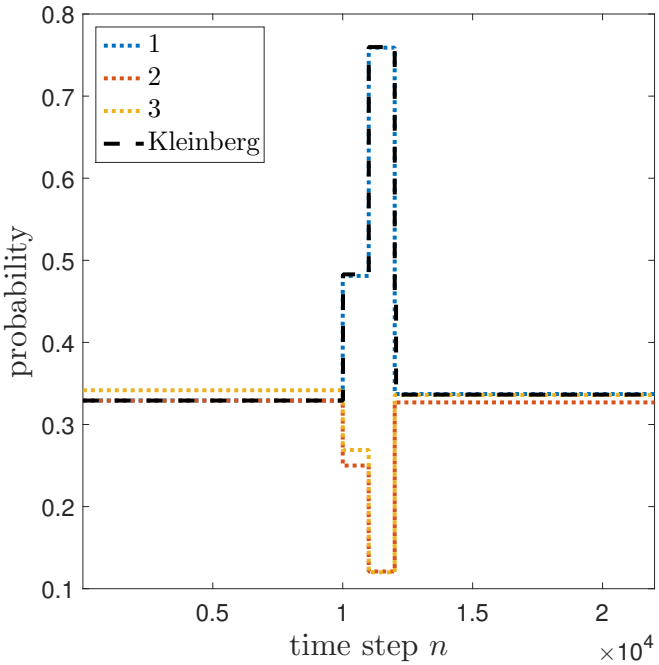

図2.11　異常区間の観測ステップ = 1000 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

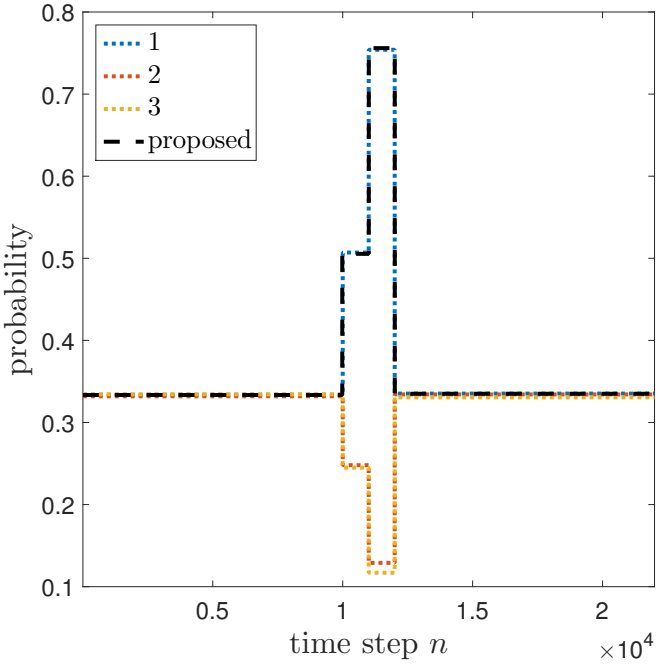

図2.12　異常区間の観測ステップ = 1000 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

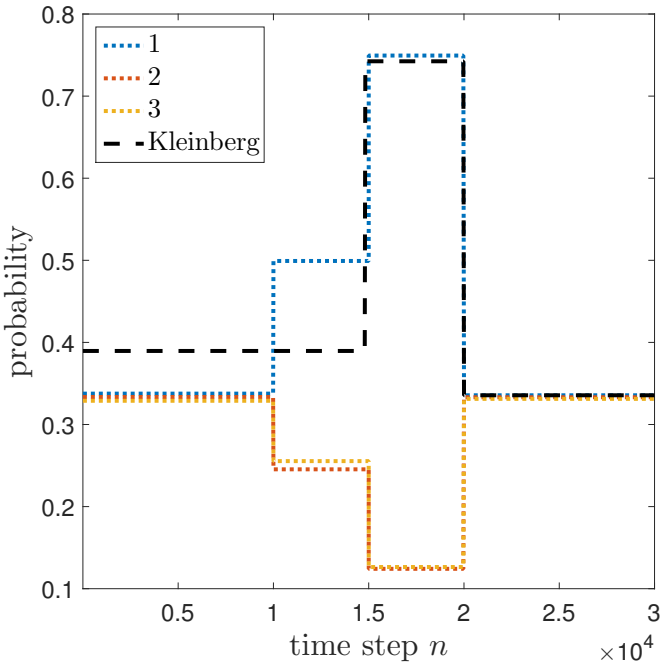

図2.13　異常区間の観測ステップ = 5000 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

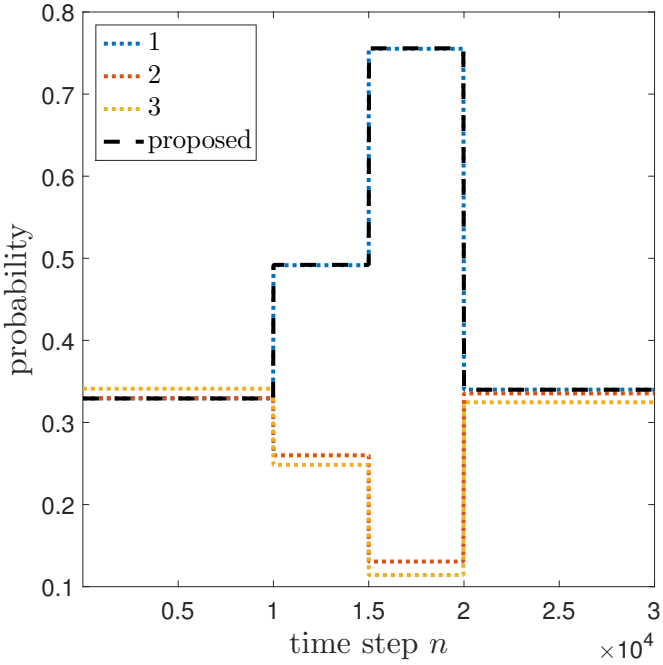

図2.14　異常区間の観測ステップ = 5000 における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

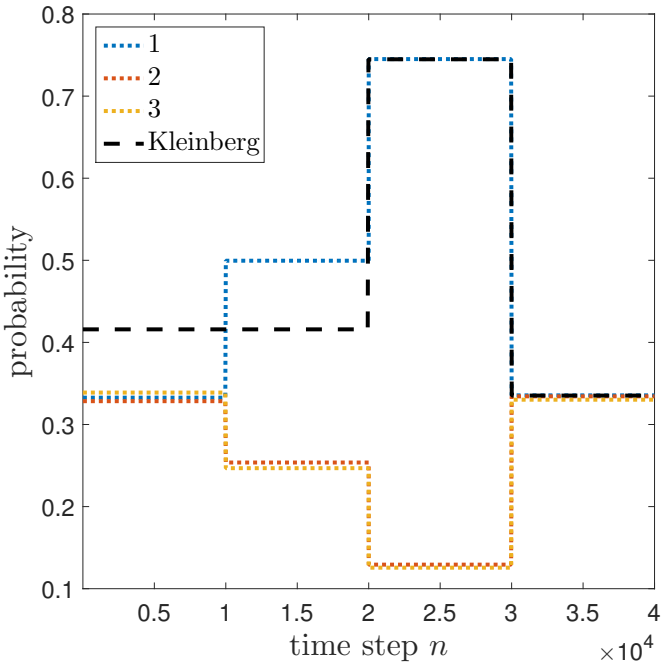

図2.15　異常区間の観測ステップ = 10000 における出現確率の絶対誤差の中央値に最も近い結果 (Kleinberg の手法)

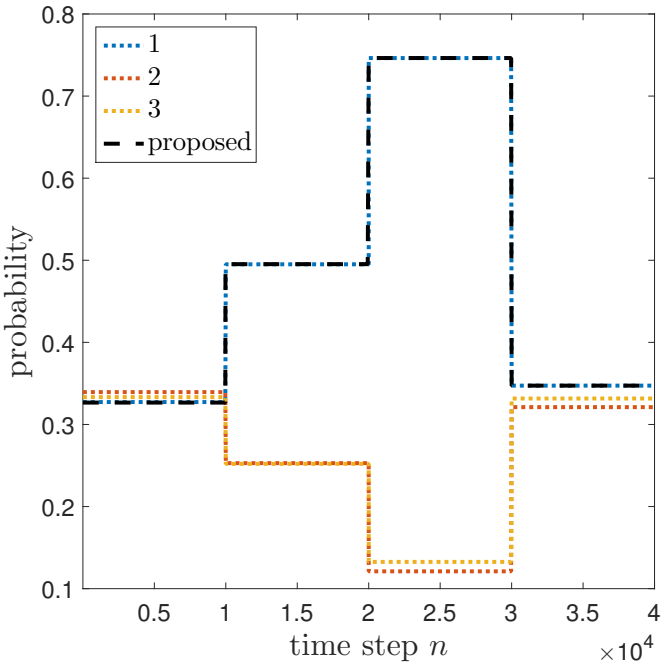

図2.16　異常区間の観測ステップ $= 10000$ における出現確率の絶対誤差の中央値に最も近い結果 (提案手法)

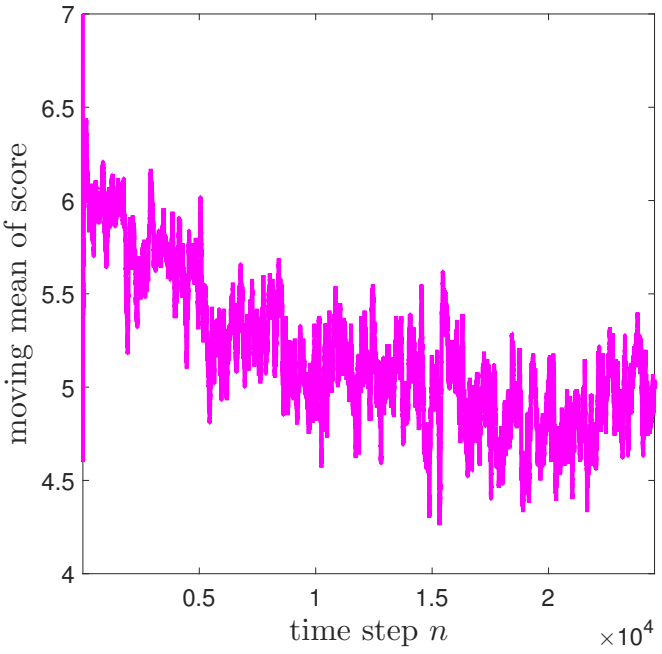

図2.17　“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” のレビュー点数の移動平均 (計算範囲100観測ステップ)

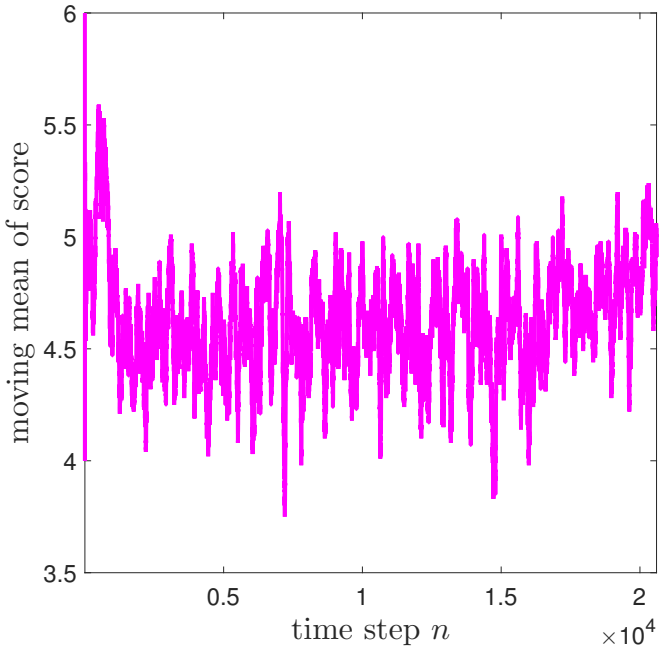

図2.18　“Conditioner Essential (Albion)” のレビュー点数の移動平均 (計算範囲100観測ステップ)

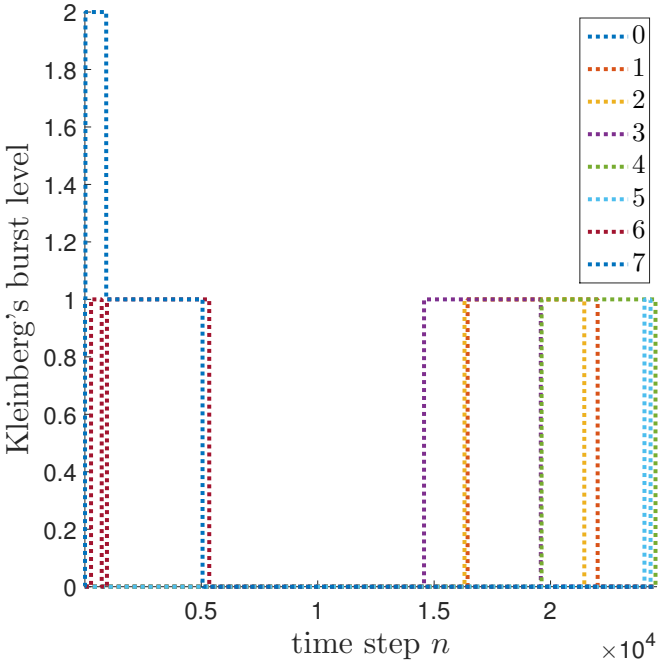

図2.19　“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” の結果 (Kleinberg の手法におけるバーストレベル)

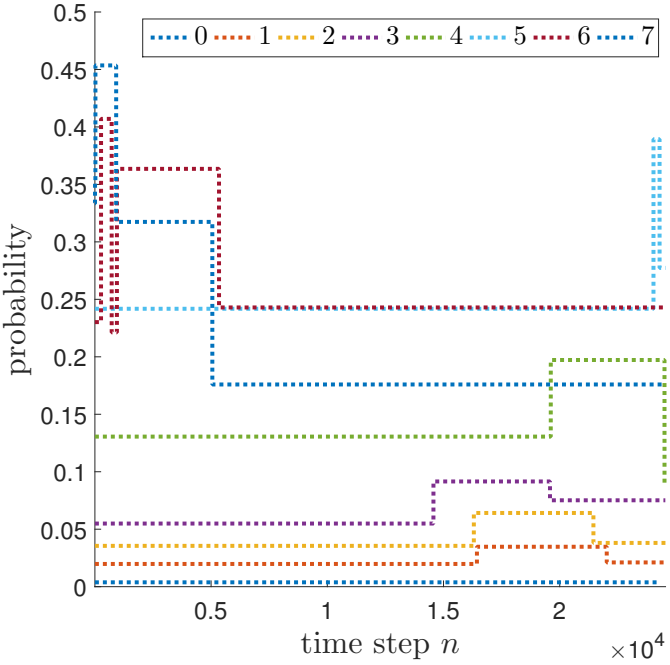

図2.20　“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” の結果 (Kleinberg の手法におけるバーストレベルを基にした出現確率)

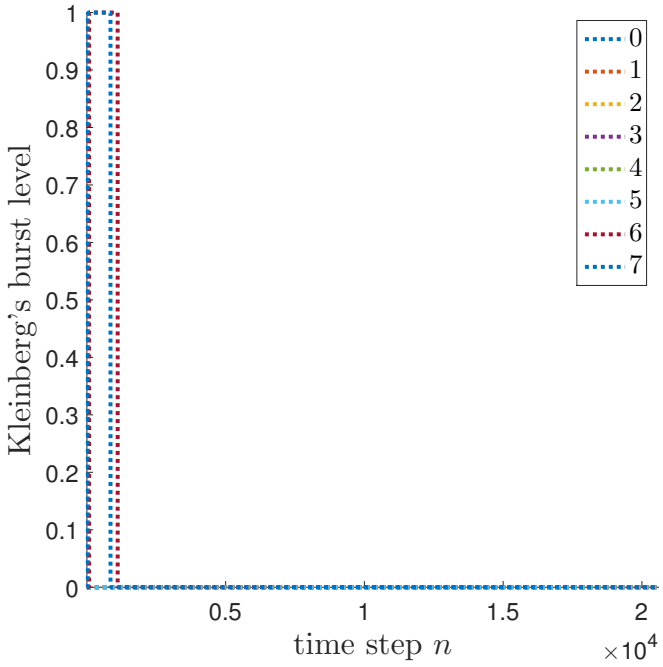

図2.21　“Conditioner Essential (Albion)” の結果 (Kleinberg の手法におけるバーストレベル)

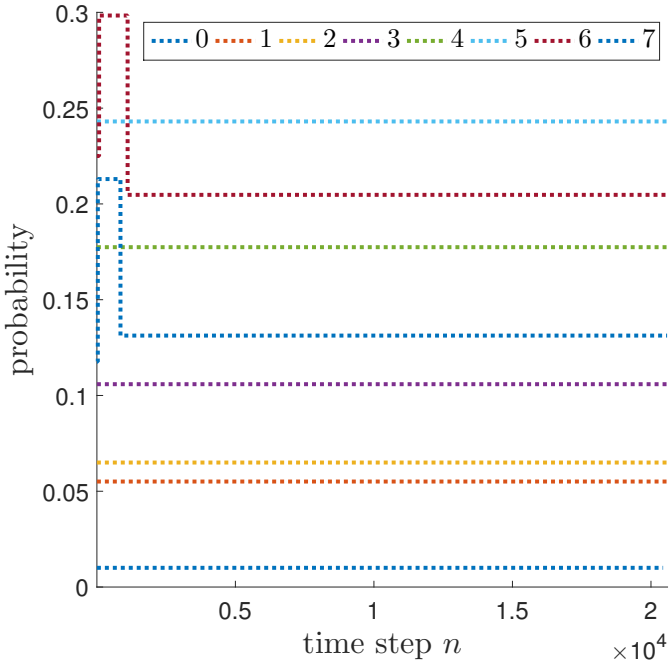

図2.22　“Conditioner Essential (Albion)” の結果 (Kleinberg の手法におけるバーストレベルを基にした出現確率推移)

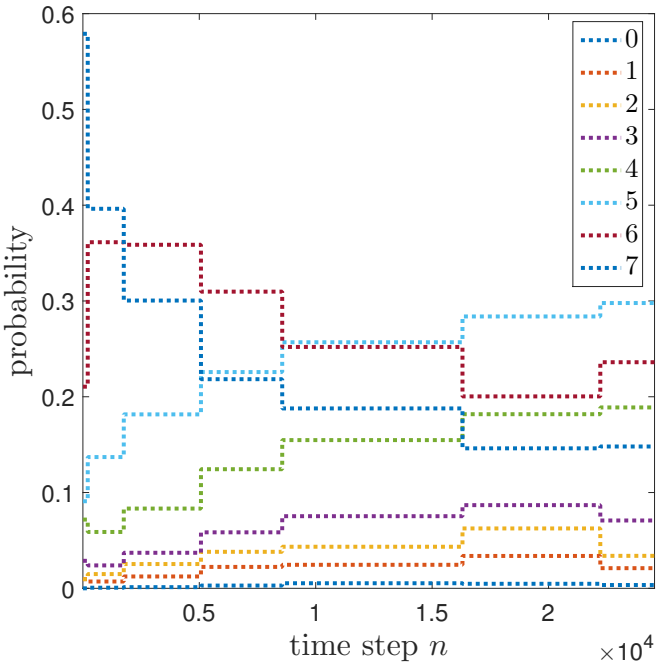

図2.23　“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” の結果 (提案手法における分割結果とそれを基にした出現確率分布)

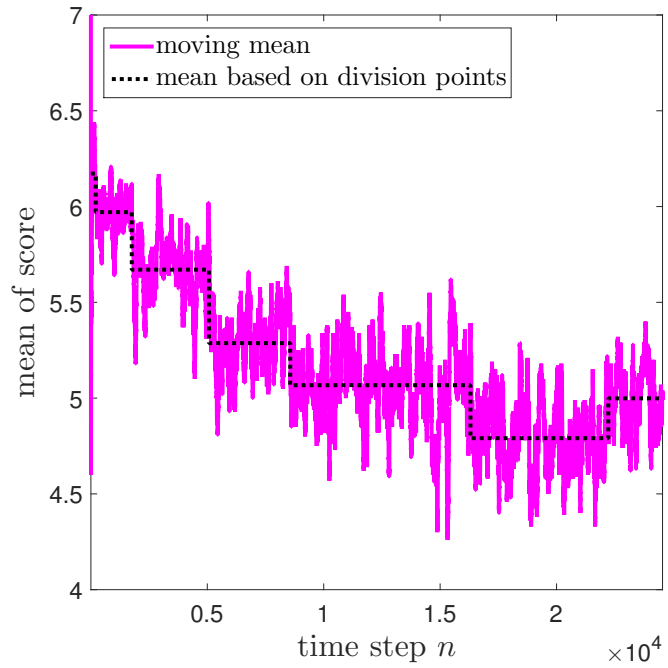

図2.24　“Oshima Tsubaki Camellia Hair Care Oil (Oshima Tsubaki)” のレビュー点数の移動平均 (計算範囲100観測ステップ) と提案手法の分割に基づく平均推移

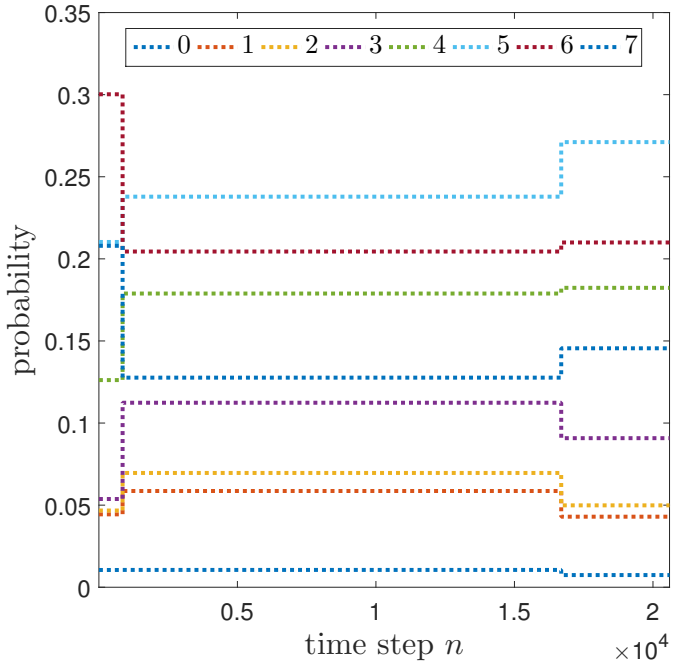

図2.25　“Conditioner Essential (Albion)” の結果 (提案手法における分割結果とそれを基にした出現確率分布)

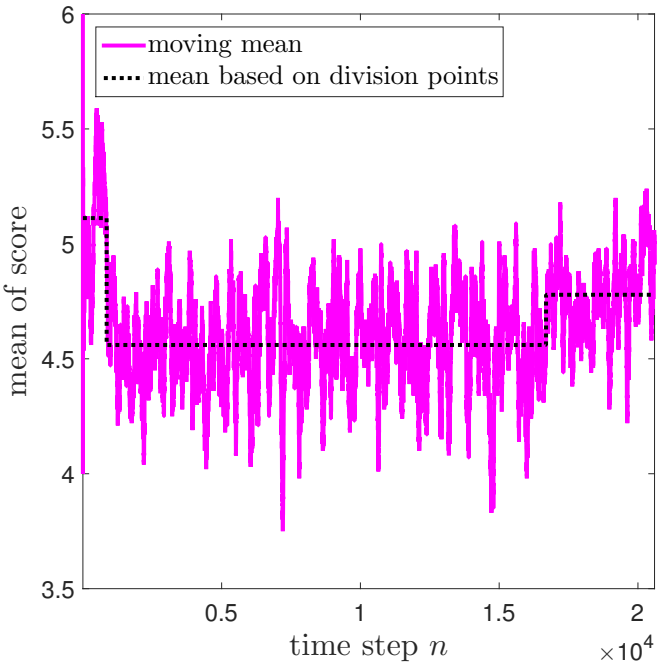

図2.26　“Conditioner Essential (Albion)” のレビュー点数の移動平均 (計算範囲100観測ステップ) と提案手法の分割に基づく平均推移

第3章

区間分割法：大規模データへの応用

この章では，第 2章で提案した区間分割法の大規模データへの応用について述べる．ここでは，大規模なレビューサイトのデータを用いて，提案した区間分割法がレビューサイト特有の問題をどこまで扱えるかを検証する．更に，高次元属性を持つようなデータへの応用についてもここで述べる．

なお，数式や解法においては，第 2章と共通の定義を使用している．

3.1　はじめに

近年，レビューサイトにおけるユーザのレビュー行動が非常に活発であり，日々大量に投稿されるレビュー情報は，購買行動を始めとしたユーザの多様な活動に影響を及ぼしている．これは，レビューサイトそのものがメディアとして，商品やサービスのプロモーションを左右する重要な情報発信源になりつつあることを示唆している．しかし，大手のレビューサイトは，レビューそのものは大量に保有しているものの，それらをユーザに情報として上手く発信できているとは限らない．何故なら，常に情報に対して受け身であるライトユーザなどは，「被レビュー数が多い人気商品」や「投稿数が多い有名ユーザのレビュー」のような情報さえ得られれば，それで満足してしまうように思えるからである．現に，Salganik らのレビューサイトに関する大規模な実験 [1] では，アイテムの質そのものよりも，多数のユーザによって生成された社会的情報の方が圧倒的に影響力が強く，ユーザ本来の嗜好すらも捻じ曲げる恐れがあるという結果が出ている．つまり，サイト利用者の大半を占めているであろうライトユーザは，本来自分が欲しているような情報よりも，レビュー数という社会的情報によって信頼性が保証されている情報を優先して得ている可能性が高いということである．これは，大量に存在するレビューのうちのほんの一部しか情報として活かされていないこと，さらに，その他の多くのレビューが「数」という情報に落とし込められてしまっていることを意味している．かと言って，大量のレビューの中から自分にとって有益な情報を的確に導き出そうとすると，多くの時間と手間を要してしまうため，社会的情報に頼るのは当然と言える．

アイテムに対する社会的情報として，広く用いられているのは平均評点である．確かに，平均評点は

アイテムに対する評価の収束結果として非常に分かりやすく，レビュー数が多ければ多いほどその信頼性が保証されるのも明確である．しかし，その平均評点に至るまで，評点の付き方が異常なまでに変化していた場合は，一概に信頼性が高いとは言い切れない．中でも代表的な例は「サクラ」や「やらせ」といった，ユーザによる意図的な評価の操作で，これらに関しては国内のニュースでも取り上げられる程問題となっている．さらに，意図的でなくとも，他のメディアから発せられた情報によって評点の付き方が変化する様は安易に観測することができ，その理由も商品の改良から流行の変遷まで多種多様である．少なからず起きているこれらのイベントは，各アイテムにおいて重要な情報であるはずだが，平均評点と直近のレビューから読み取ることは難儀である．それどころか，先立った情報が無ければ，地道に一つずつレビューを辿っていっても誰もが気付けるわけではない．よって，膨大なレビューの中からこれらのイベントを検出し，アイテムに関する情報として利用できるようにすることは，Web情報学において非常に重要な研究と言える．本章では，第 2章で提案した区間分割法を用いて，現実の大規模データにおけるイベントの検出を試みる．

3.2　評価実験

3.2.1　データセット

実験で使用したデータセットは，第 2章でも用いた @cosme[*1]のレビューデータセットである．本章ではレビューデータの全体に提案手法を適応するため，ここで一度 @cosme とデータセットの詳細について述べる．@cosmeは，株式会社アイスタイル[*2]が運営する日本最大級の化粧品レビューサイトであり，1999年12月にサービスが開始された．このデータセットは，2013年8月に@cosmeをクロールして取得したものであり，443853ユーザ，6173アイテム，4376241レビューを有する．なお，クロール対象は観測ステップ数（被レビュー数） N が200以上のアイテムとした．レビューの評点は 0 から 7 の整数値を取りうるため，実験時には 1 ずつ加算して 1 から 8 ($J = 8$)として扱う．図 3.1にレビュー評点の分布を示す.

ここで，このデータセットにおけるレビューは，人間の行動特性に基づいて投稿されたものかどうかを検証する．人間の行動特性として，手紙が手元に届いてから返信するまでの時間間隔の確率分布は $f(x) = x^{-\alpha}$,$(\alpha = 3/2)$ の冪乗則に近似することが Oliveira と Barabasi によって既に報告されている [22] ため，検証にはユーザのレビュー投稿間隔集合を用いる．

レビューデータに含まれるユーザ u とユーザ集合 U を

$$u \in U = \{u_1, \cdots, u_M\}, \tag{3.1}$$

と定義する．このとき，ユーザ u がレビューを投稿した時刻をτ_u とし，時刻の新しいものから

[*1] http://www.cosme.net
[*2] http://www.istyle.co.jp/

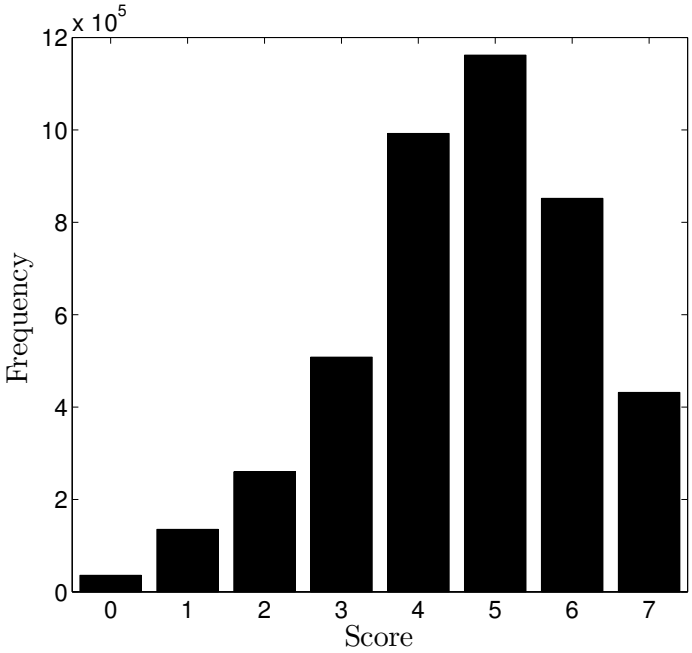

図3.1　レビュー評点の分布

$\tau_{u,1}, \tau_{u,2}, \cdots$ としていくと，ユーザ u のレビュー時刻集合 $I(u)$ は

$$I(u) = \{\tau_{u,1}, \cdots, \tau_{u,|I(u)|}\}, \tag{3.2}$$

となり，ユーザ u のレビュー投稿間隔集合 $\Delta I(u)$ は，

$$\Delta I(u) = \{\Delta\tau_{u,i} = \tau_{u,i} - \tau_{u,i-1} : i = 2, \cdots, |I(u)|\}, \tag{3.3}$$

となる．よって，ユーザのレビュー投稿間隔集合 I は，

$$I = \bigcup_{u \in U} \Delta I(u), \tag{3.4}$$

で書き表すことができる．

図 3.2は，データセットから求めたレビュー投稿間隔集合Iの$\Delta\tau_{u,i}$を日数に変換したときの確率分布であり，破線は$f(\Delta\tau_{u,i}) = \Delta\tau_{u,i}^{-\alpha}$,$(\alpha = 3/2)$を示す．図より，ユーザの投稿間隔の確率分布は，$\alpha = 3/2$ の冪乗則に近い分布となっているため，このデータセットのレビューには，人間の行動特性が少なからず反映されていると言える．

3.2.2　実験結果

データセットのアイテム毎で得られた分割点集合 $\mathcal{C}_K$ に関する結果を示す．今回，実験対象の観測ステップ数 N が 200 から約2万と幅広く存在しているため，貪欲法 (A1) における χ^2 の危険率は

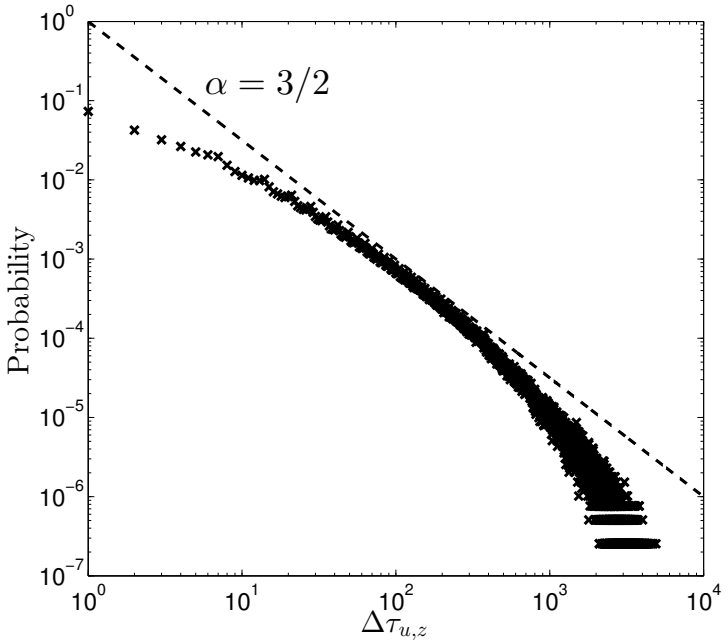

図3.2　ユーザのレビュー投稿間隔分布

$p = 0.005$（検出感度は $p = 0.0001$ よりも強め）に設定した．ここでは，区間分割法の基本的な挙動の確認と，局所探索法 (A2) の直接的な評価をするべく，A1 と A2 を逐次的に行う従来解法の結果について述べる．

図 3.3，3.4，3.5 に，K の度数分布，k と $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ のプロット，K と観測ステップ数 N のプロットをそれぞれ示す．また，表 3.1 には $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ の上位10アイテムを示す．図 3.3,3.4より，K の度数は冪乗則に従っており，k と $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ の関係も似たような性質を持っていることが分かる．また，図 3.5 より，分割点数 K はアイテムの観測ステップ数 N に大きく依存していることが分かる．なお，この実験で得られた分割点の総数は8119である．

さらに，図 3.6と 3.7に，K ごとの解の改善値平均と，K ごとの計算量の増加比率平均を示す．図 3.6 の縦軸は，局所探索法 (A2) のプロセスにおいて改善された目的関数値の平均を表しており，また，図 3.7 の縦軸は，局所探索法によって増加した計算量の比率（全体の計算量/貪欲法のみの計算量）の平均を表している．図 3.6 より，K の増加と共に改善値が増加することは明確であり，また，図 3.7 より，全体の計算量は貪欲法の高々3倍程度にしかならないことが分かる．

3.2.3　分割点分析

ここでは，$\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ の上位アイテムに対し，詳細な分析を行う．因みに，実際の上位 10 アイテムは表 3.1 の通りであるが，観測ステップ数が多い例については第 2章で既に紹介している

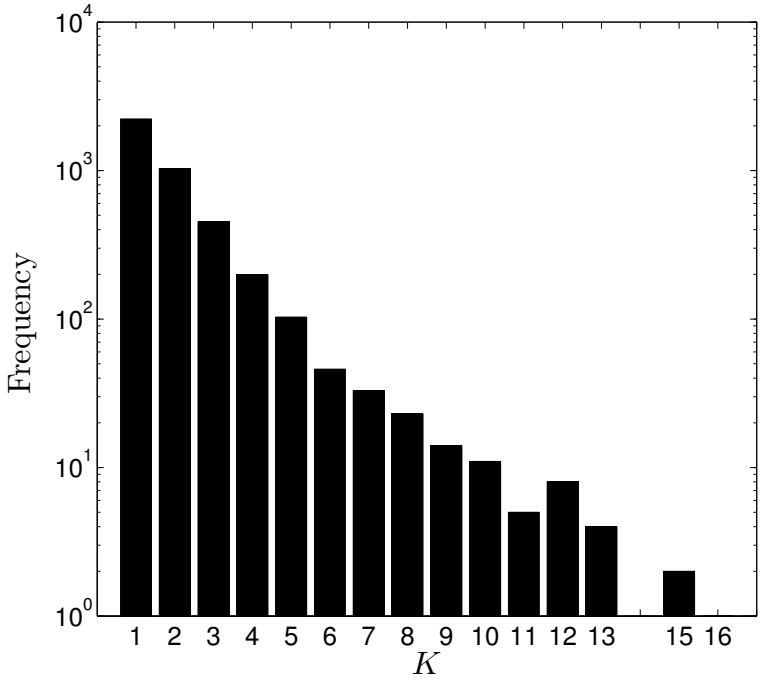

図3.3　K の度数分布

表3.1　$\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ 上位10アイテム

No.	Item Name	k	Value	N
1	大島椿(ツバキ油)	1	665.35	24552
2	オートマティックライナー	1	586.70	7730
3	シャドーカスタマイズ	1	396.15	11127
4	NU ソワンオレオリラックス	1	286.92	9853
5	みつばちマーチ	1	277.13	12380
6	シアトリカルパウダー	1	271.77	9590
7	エクスプレスケアトータルクリーン	1	241.82	2196
8	ビューティーコスメマスカラ	1	203.34	8382
9	資生堂眉墨鉛筆	1	194.65	7056
10	ロングウェアジェルアイライナー	1	192.04	7759

ため，今回は観測ステップ数が少ない場合の分析例として $N \leq 500$ のアイテムを紹介する．以下に，$N \leq 500$ という条件において，$\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ が最も上位であった2アイテムの分割点分析について述べる．

1つ目はカットコットン(無印良品)というアイテムで，全 $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ 中 $k = 1$ が78位，$k = 2$ が5626位である．図 3.8 は，従来解法による分割結果と，それを基にした評点の出現確率分布を

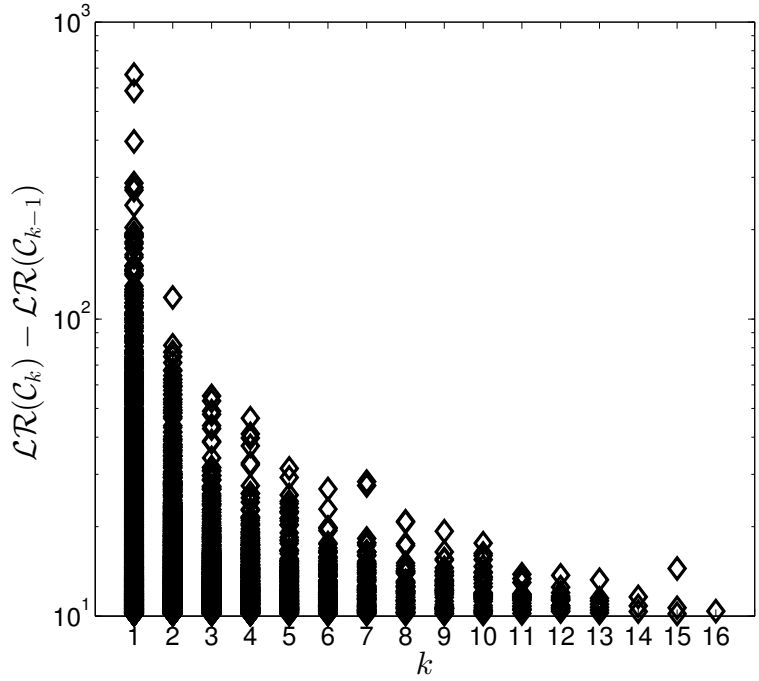

図3.4　k と $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ のプロット

示したものであり，図 3.9 は，レビュー点数の移動平均(計算範囲10観測ステップ) と，分割結果に基づく平均評点の推移を示したものである．両図より，このアイテムは，T_2 以降高評価のレビューで安定していることが分かる．実際に T_2 前後 10 レビュー程度を確認すると，アイテムそのものの改良が大きく影響していることが分かった．より具体的には，「毛羽立ちが抑えられている」や「肌触りが良くなった」という意見が多く見られ，枚数を増やしたのに価格を据え置きにしていることについても高い評価を得ていた．この分割点には，情報に敏感なユーザ (early adopters) が大きく関わっていることが分かる．次いで T_1 前後 10 レビュー程度を確認すると，初期のレビュー評価が低すぎると感じたユーザによる対抗レビューが T_1 以降多く書かれており，評価の回復傾向が見られた．

2つ目は魔法の泡立てネット(モンクレール)というアイテムで，全 $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ 中 $k = 1$ が81位，$k = 2$ が1410位，$k = 3$ が3132位である．先程と同様，図 3.10 に従来解法による分割結果と分割に基づいた評点の出現確率分布を，図 3.11 にレビュー点数の移動平均と分割に基づいた平均評点推移を示す．このアイテムの場合，T_2 から評価の低迷が見られるので，実際に T_2 前後 10 レビュー程度を確認すると，他社が頑丈なものを低価格で販売し始めたことが影響していることが分かった．さらにレビューを過去に遡って見ていくと，発売当初は泡立てネットそのものが市場に存在していなかったため，商品の目新しさによって高評価のレビューが集まったことも分かった．類似商品が出現し始めた T_2 以降は，それらと比較して低評価としているレビューが多く，発売当初とのギャップが大きい．先

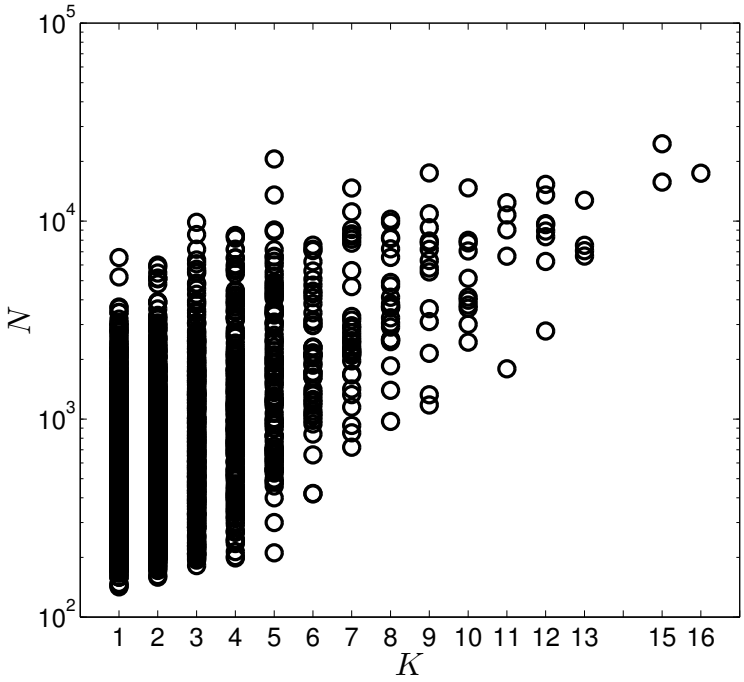

図3.5　K と N のプロット

程と同様，この分割点も，情報の持ち込みが早いユーザ (early adopters) によって引き起こされたものと見られる．一方，T_3 前後 10 レビュー程度とそれ以降では，「長期間使用した」とレビューに明記しているユーザの高評価レビューが出現し始め，評価にバラつきが生じている．T_1 から T_2 の間は，辛口のレビュアによる厳しい批評も登場するが，レビュー傾向は T_1 以前と同様，高評価寄りであった．

ここで，提案法による分割結果は，レビューの投稿時刻や投稿間隔といった時間的な密度を一切考慮していないことに注意されたい．言い換えれば，これらの分割結果は，アイテムがいつ注目されたかという「話題性」を一切考慮していないということであり，逆に考えれば，「話題性」に左右されていないということである．

3.3　計算量と解品質

計算量と解品質の関係をさらに追求するべく，従来解法と提案解法を比較する．

3.3.1　解法の比較

以下に，@cosme データセットにおける解法の比較結果を述べる．図 3.12，3.13，3.14 はそれぞれ，K の度数分布，K ごとの解の改善値平均，K ごとの計算量の増加比率平均についての比較を示したも

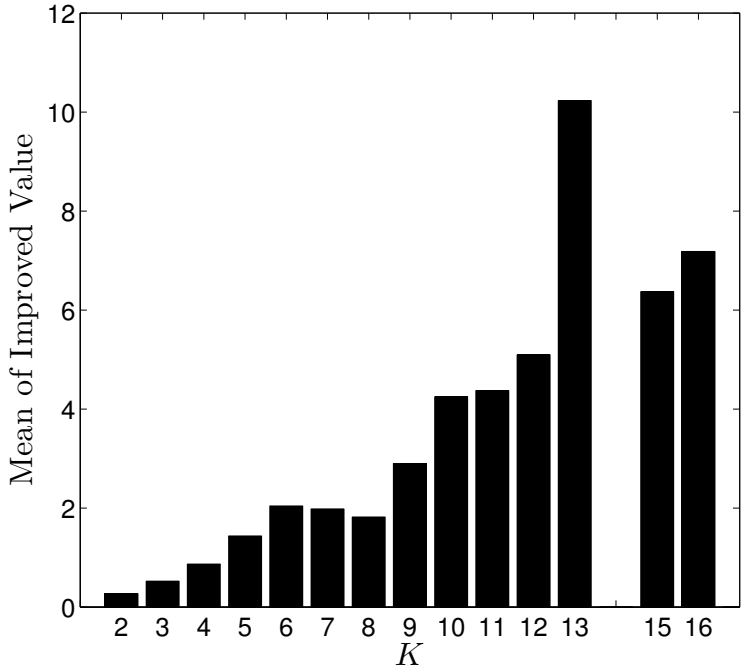

図3.6　K ごとの局所探索法による解の改善値平均

のであり，黒棒が従来解法の結果，白棒が提案解法の結果である．

図 3.12 より，提案解法で得られる分割点数の期待値は，従来解法よりも低いことがわかる．分割点の総数で見ても，従来解法が 8119 だったのに対し，提案解法は 7930 であった．また，図 3.13 より，提案解法は従来解法に比べて解の改善精度が高い傾向にあるので，提案解法は高い解品質が期待できる．一方，図 3.14 より，提案解法の計算量は約 K 倍に膨れ上がっているため，従来解法の方が，さらに大規模なデータでも高速に解けることが期待できる．

3.4　高次元への応用

補足実験として，扱う属性 J が高次元の場合の応用について述べる．

3.4.1　データセット

ここで用いたデータセットは，第7回静岡おまちバル [*1]のために設置したおよそ 100 箇所の Free Wi-Fi Spot のアクセスログデータである．ユーザ ID はデバイスの MAC アドレスで分類し，アクセス先サーバ ID はアクセス先の問い合わせ回答にオーソリティを持つ DNS で分類をしている．データ

[*1] http://www.omachibar.com

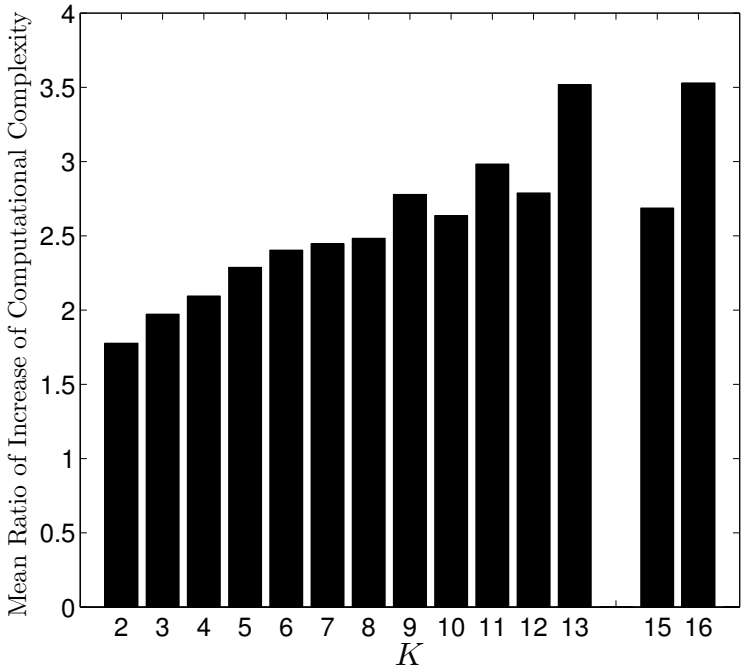

図3.7　K ごとの局所探索法による計算量の増加比率平均

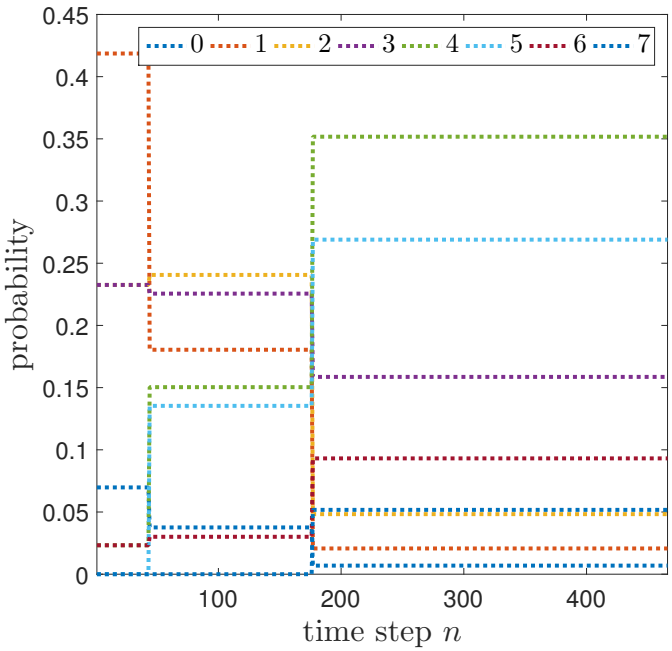

図3.8　カットコットンの結果(従来解法における分割結果とそれを基にした出現確率分布)

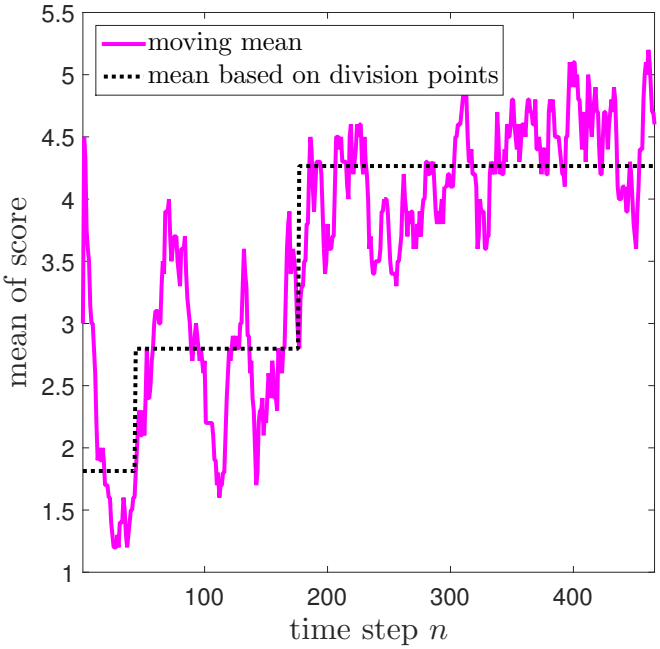

図3.9　カットコットンのレビュー点数の移動平均(計算範囲10観測ステップ) と従来解法の分割に基づく平均推移

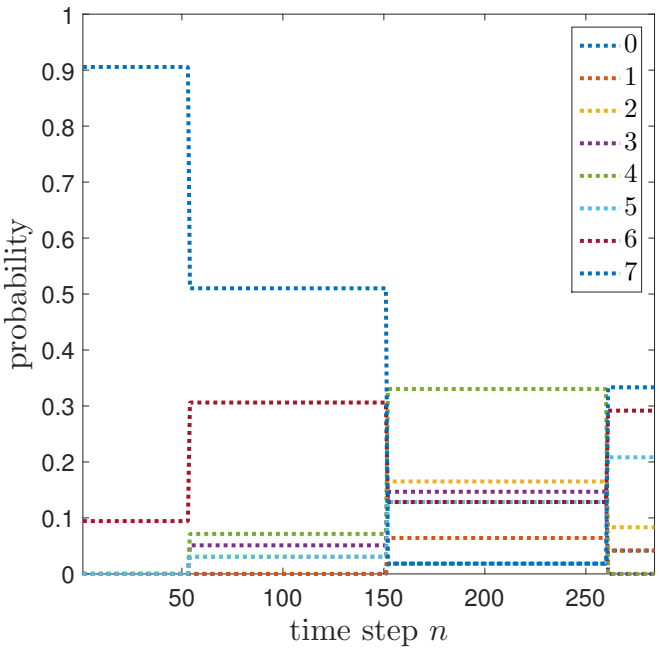

図3.10　魔法の泡立てネットの結果(従来解法における分割結果とそれを基にした出現確率分布)

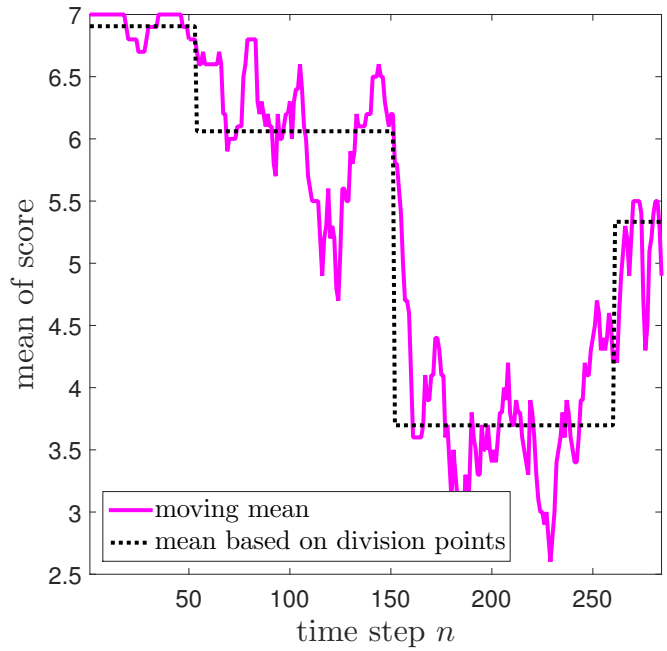

図3.11　魔法の泡立てネットのレビュー点数の移動平均(計算範囲10観測ステップ) と従来解法の分割に基づく平均推移

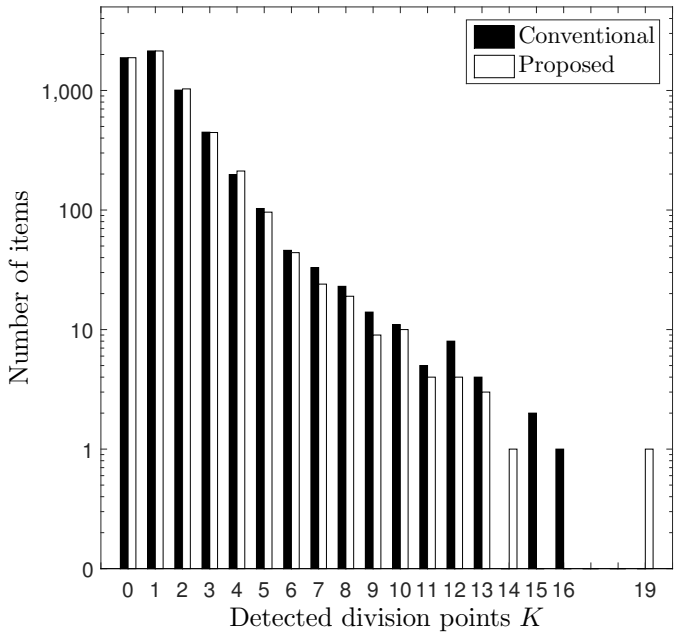

図3.12　K の度数分布の比較

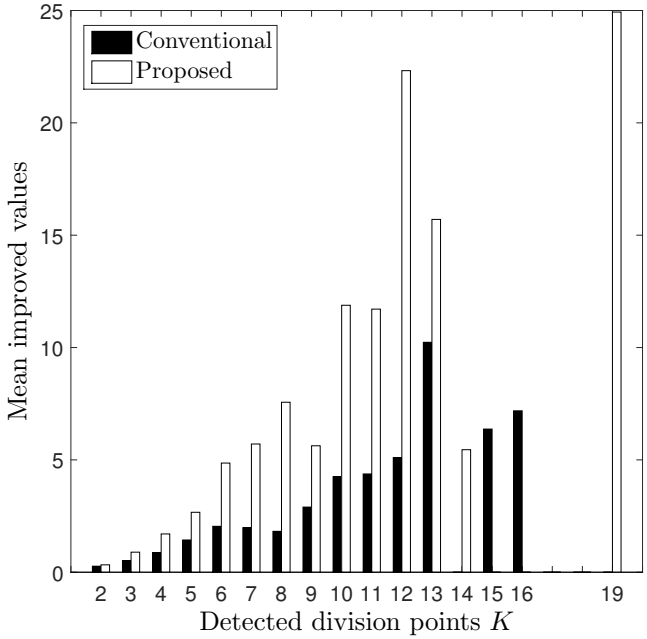

図3.13　K ごとの解の改善値平均の比較

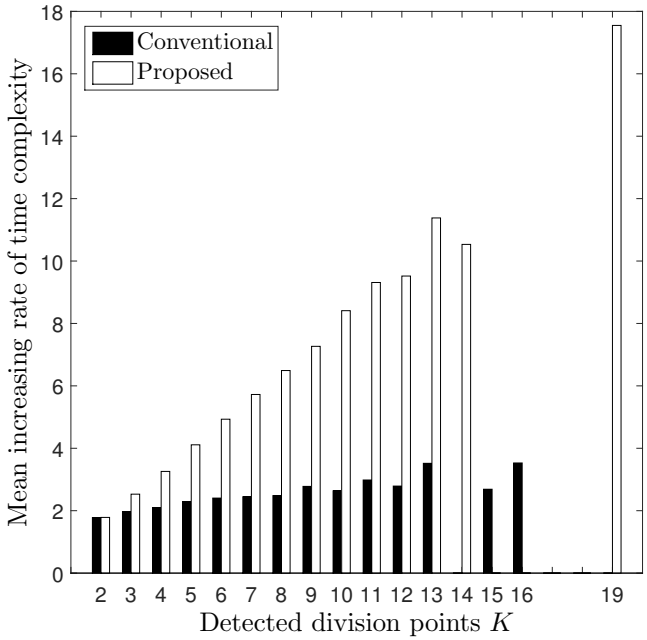

図3.14　K ごとの計算量の増加比率平均の比較

セットの対象時刻は 11/08/2014 09:08:35 から同日の 22:01:09 であり，238 ユーザ，299 アクセス先サーバを有する．このデータに対し，ユーザ毎のアクセスアクティビティについて区間分割法を適応しようとすると，$J = 299$ の高次元属性の区間分割問題となる．ここでの検出感度（χ^2 の危険率）は $p = 0.001$ とした．

3.4.2　実験結果

ここでは，貪欲法 (A1) のみで解く単純解法 (*Simple*)，従来解法，提案解法の3手法で比較を行う．実験における各手法の解品質の比較を表 3.2 に，計算時間 (Intel(R) Xeon(R) CPU X5690 @ 3.47GHz) の比較を表 3.3 にそれぞれ示す．表より，提案解法は他の2手法よりも計算時間を要しているが，実用的な時間で解くことができており，解品質においては他の2手法よりも優れている．

また，提案解法における代表結果として，ユーザ ID 13 のアクセスログと分割点を図 3.15 に，各区間の主要 DNS アクセス確率分布を図 3.16 にそれぞれ示す．図より，本手法の分割点は，アクセス分布の変化を基に，アクセスログを有意に分割することができていることが見て取れる．例えば，$\mathcal{D}_1$，$\mathcal{D}_3$，$\mathcal{D}_5$は，似たようなアクセス分布となっており，特定のドメイン (DNS 14) へのアクセスが多いことから，仕事関連の作業をしていることが推察される．更に，$\mathcal{D}_2$，$\mathcal{D}_6$は，大手ポータルサイト (DNS 10) へのアクセスが多いことから，ウェブブラウザで検索作業をしていることが分かる．また，$\mathcal{D}_4$，$\mathcal{D}_7$は，他の区間には出現しないアクセス先サーバへのアクセスが多いため，デバイスの通信状態が他の区間とは明確に異なっている．

表3.2　解品質の比較

	Sum of $\mathcal{LR}(\mathcal{C}_K)$	Rate
Proposed	34805.6837	101.01%
Conventional	34771.0212	100.91%
Simple (A1 alone)	34459.0665	100.00%

表3.3　計算時間の比較

	Calculation time	Rate
Proposed	104.99 sec	1418.78%
Conventional	24.13 sec	326.08%
Simple (A1 alone)	7.40 sec	100.00%

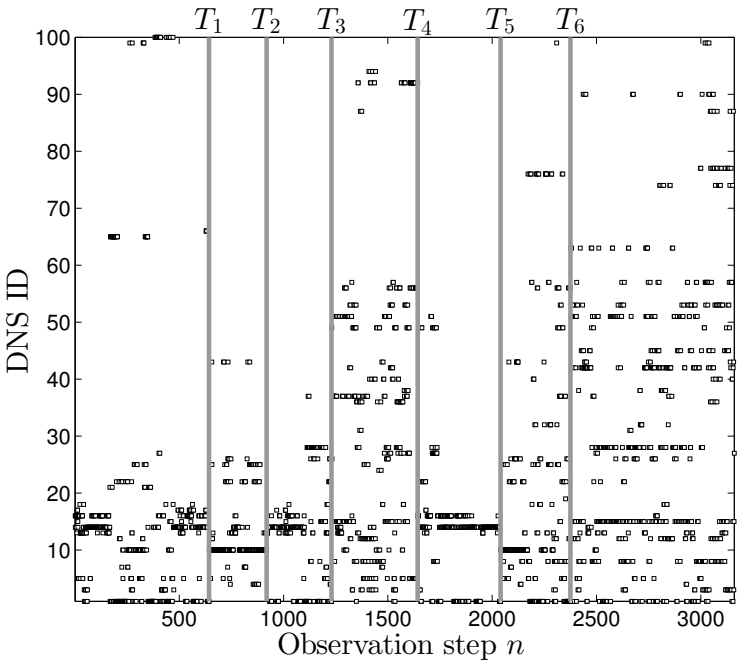

図3.15　ユーザ ID 13 のアクセスログと分割点

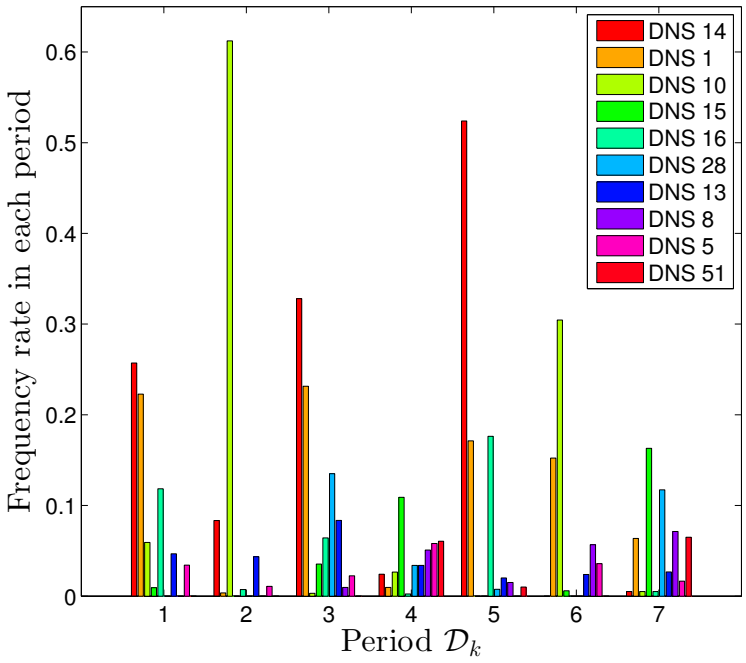

図3.16　ユーザ ID 13 の主要 DNS アクセス確率分布

3.5　おわりに

第 2章で提案した区間分割法の大規模データへの応用として，レビューデータを用いた実験を行った．実験では，区間分割法がレビューの分析支援として十分有用であることを示した．また，今回提案した解法における解品質と計算量の関係についても深く追求した．さらに，区間分割法は高次元属性の時系列データにおいても高速で解を求めることができ，高次元の結果についても，分析支援の有効性を示した．

第4章

区間分割法：信頼区間と異常区間

この章では，第 2章で提案した区間分割法の結果の利用方法について述べる．分割した区間において，特に区間の異常度が高いものは異常区間として評価の対象外としたり，直近の分割区間だけを信頼区間として評価対象としたりすることがここにおける目的である．ここでは，区間を z-score として扱う方法と，それを利用したランキング手法について提案する．

なお，数式や解法においては，第 2章，第 3章と共通の定義を使用している．

4.1 はじめに

レビューサイトにおけるレビュー対象アイテムのランキングは，殆どの場合，レビュー投稿数やレビュー平均評点といったナイーブなソーシャル情報や，公表されていないサイト独自の手法，とりわけユーザ依存の信頼を用いたものによって生成されている．確かに，ランキングの秩序を守るためには，独自の方法で信頼性の低いユーザのレビュー情報を淘汰し，その手法を公表しないというのも重要であるが，その不透明性故に，ユーザがランキングの信頼性を懸念する可能性も大いにある．更に，ユーザに提供するナイーブなソーシャル情報の項目数を増やすと，個々の意思決定に多大な影響を与え，市場の不平等性を大いに増加させることが Salganik らの大規模実験 [1]によって既に分かっている．よって，既存のランキングに対する代替案の一つとして，ユーザ依存の信頼ではなくアイテム依存の信頼を利用し，且つナイーブなソーシャル情報のみに依存しないような，統計モデルに基づくランキングの構築が考えられる．

したがって，本章では，既に第 3章でレビューデータにおける有効性を示した区間分割法の結果を利用して，アイテム毎の評価方法を展開する．アイテム依存の信頼性を考慮するためには，各区間の異常度を定式化する必要があるため，それについてもこの章で扱う．本章における信頼とはアイテム依存のものであるため，推薦システムの研究 [23, 24]で頻繁に用いられているような，ユーザ依存の信頼とは異なることに注意されたい．

4.2　ランキング手法

レビュー評点 j が与えられる確率は多項分布であることを仮定しているので，レビューデータセットにおける平均と標準偏差はそれぞれ

$$\mu = \sum_{j\in\mathcal{J}} j\bar{p}_j, \tag{4.1}$$

$$\sigma = \sqrt{\sum_{j\in\mathcal{J}} (j-\mu)^2\bar{p}_j}, \tag{4.2}$$

のように算出される．ここで，$\bar{p}_j$ はデータセット全体における平均 $E(\hat{p}_j)$ である．各レビュー評点が，評点分布 $\bar{p}_j$ に従って独立に与えられたと仮定すると，$\mathcal{Q}$ 個のレビュー $\mathcal{S} = \{s_1, \cdots, s_\mathcal{Q}\}$ が投稿されたときの標準偏差，すなわち評点の期待値からの偏差の二乗平均平方根($RMSE$)は以下となる．なお，ここで $\langle\cdot\rangle$ は期待値を表す．

$$\begin{aligned}
RMSE &= \sqrt{\sum_{s_1\in\mathcal{J}}\cdots\sum_{s_\mathcal{Q}\in\mathcal{J}}\left(\mu - \frac{1}{\mathcal{Q}}\sum_{q=1}^{\mathcal{Q}} s_q\right)^2 \prod_{q=1}^{\mathcal{Q}} p(s_q)} \\
&= \sqrt{\left\langle\left(\mu - \frac{1}{\mathcal{Q}}\sum_{q=1}^{\mathcal{Q}} s_q\right)^2\right\rangle} \\
&= \sqrt{\frac{1}{\mathcal{Q}^2}\left\langle\left(\sum_{q=1}^{\mathcal{Q}}(s_q-\mu)\right)^2\right\rangle} \\
&= \sqrt{\frac{1}{\mathcal{Q}^2}\left\langle\sum_{q=1}^{\mathcal{Q}}(s_q-\mu)^2 + \sum_{x\in\mathcal{Q}}\sum_{q\in\mathcal{Q},q\neq x}(s_x-\mu)(s_q-\mu)\right\rangle}.
\end{aligned} \tag{4.3}$$

ここで，$\langle(s_q-\mu)^2\rangle$は定義によるところの分散$\sigma^2$であり，$\langle s_q\rangle = \mu$なので，

$$\begin{aligned}
RMSE &= \sqrt{\frac{1}{\mathcal{Q}^2}\sum_{q=1}^{\mathcal{Q}}\sigma^2} \\
&= \sqrt{\frac{\sigma^2}{\mathcal{Q}}} \\
&= \frac{\sigma}{\sqrt{\mathcal{Q}}}.
\end{aligned} \tag{4.4}$$

よって，あるアイテムの時刻区間 $\mathcal{D}_k$ における平均評点の z-score は，以下のように考えることができる．

$$z(\mathcal{D}_k;\hat{\boldsymbol{p}}_k,\mathcal{C}_K) = \frac{\mu(\hat{\boldsymbol{p}}_k)-\mu}{\sigma/\sqrt{|\mathcal{D}_k|}},\ \mu(\hat{\boldsymbol{p}}_k) = \sum_{j\in\mathcal{J}} j\hat{p}_{k,j}. \tag{4.5}$$

明らかに，この $z(\mathcal{D}_k;\hat{\boldsymbol{p}}_k,\mathcal{C}_K)$ が大きければ（小さければ），時刻区間 $\mathcal{D}_k$ における評価が有意に高い（低い）とみなすことができる．更に，この z-score と区間の長さを用いて，区間の異常値を

$\sqrt{|z(\mathcal{D}_k;\hat{\boldsymbol{p}}_k,\mathcal{C}_K)|/|\mathcal{D}_k|}$ として算出することもできる．

今回，区間分割を考慮しない評価値 $z(\mathcal{D}_1;\hat{\boldsymbol{p}}_1,\mathcal{C}_0)$ を *basic* (基本多項分布法)，異常値の基準 A を超える異常区間を除いた評価値 $z(\mathcal{D}';\hat{\boldsymbol{p}'},\mathcal{C}_K)$ ($\mathcal{D}' = \{\mathcal{D}_k \in \mathcal{D} | \sqrt{|z(\mathcal{D}_k;\hat{\boldsymbol{p}}_k,\mathcal{C}_K)|/|\mathcal{D}_k|} \leq A\}$, $\hat{p'}_j = \sum_{n\in\mathcal{D}'} s_{n,j}/|\mathcal{D}'|$) を *basic'* (異常区間除去法), 最も新しい時刻区間を用いた評価値 $z(\mathcal{D}_{K+1};\hat{\boldsymbol{p}}_{K+1},\mathcal{C}_K)$ を *latest* (信頼区間採用法)とする．

4.3　データセット

今回使用するデータセットは，TripAdvisor [*1]に登録されている，日本の観光スポットのレビューデータである．取得時期は2014年10月，アイテム数（観光地域単位）は1522，総レビュー数は323868，レビュー評点は1から5の整数値($J = 5$)となっている．

4.4　実験結果

4.4.1　計算量と解品質

図 4.1，4.2，4.3に，単純解法(A1 alone)，従来解法，提案解法のそれぞれの計算時間，解品質，分割点数を示す．なお，図の横軸は A1 の終了条件である χ^2 検定の危険率 p であり，ここでの解品質は，全アイテムにおける $\mathcal{LR}(\mathcal{C}_K)$ の総和を K の総和で割ったものである．図より，提案解法は他の2手法と比べて計算時間はかかっているが，解品質においては優秀であり，分割点数も少ないため，無駄な分割を避けていることが分かる．

これ以降は，χ^2 検定の危険率を $p = 0.005$ としたときの 提案解法の実験結果について述べる．評価値の比較のためには，少なくとも1つの分割点が必要となるため，対象となったアイテム数は今回 167 である．また，図 4.4 に示す異常値の分布より，異常区間となる基準は $A = 0.20$ とした．

4.4.2　評価値の分布

図 4.5，4.6，4.7にbasic (基本多項分布法)の分布，basic' (異常区間除去法)の分布，latest (信頼区間採用法)の分布をそれぞれ示す．横軸は評価値，縦軸はレビュー数，色は全区間における平均評点を表す．なお，basic' においては，$|\mathcal{D}'| = 0$ となったもの，すなわち全区間が異常区間としてみなされたアイテムは表示されていない．図より，basic は平均評点に準拠した評価値となっているが，basic' と latest はその限りではないことが分かる．更に，basic' は basic と比べて評価値の分散が多少収まっており，latest はレビュー数に関係なく評価値が散らばっていることが見て取れる．

[*1] http://www.tripadvisor.com/

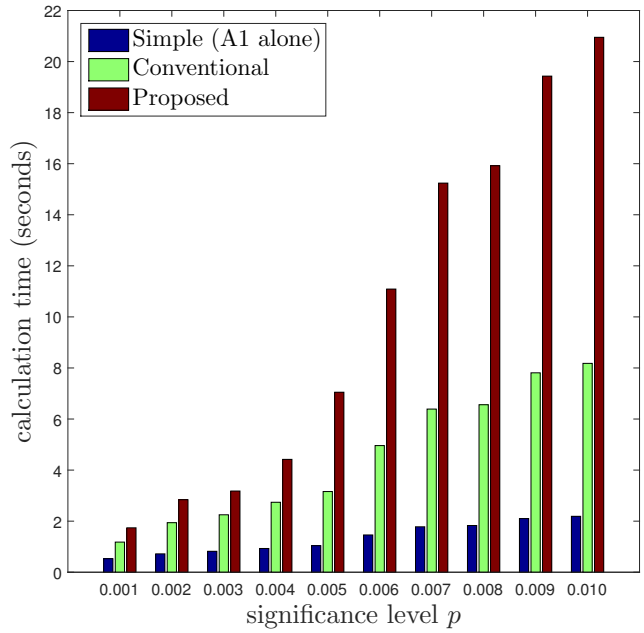

図4.1　各解法による計算時間

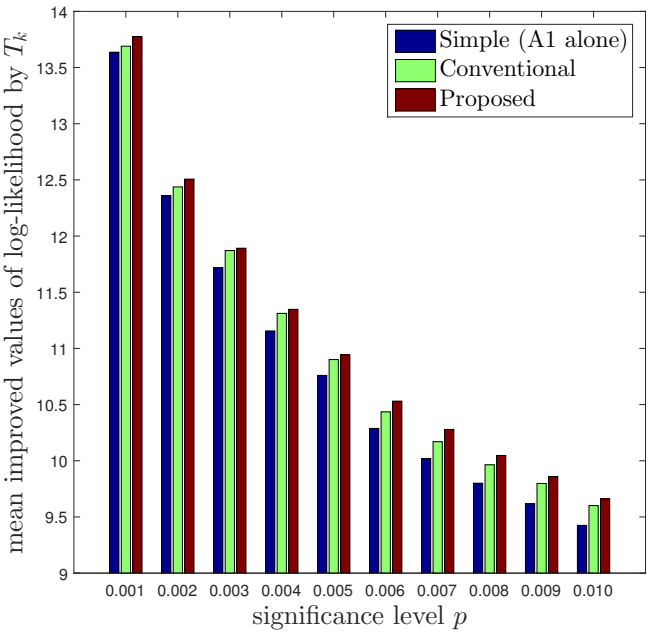

図4.2　各解法による解品質（全アイテムにおける $\mathcal{LR}(\mathcal{C}_K)$ の総和を K の総和で割ったもの）

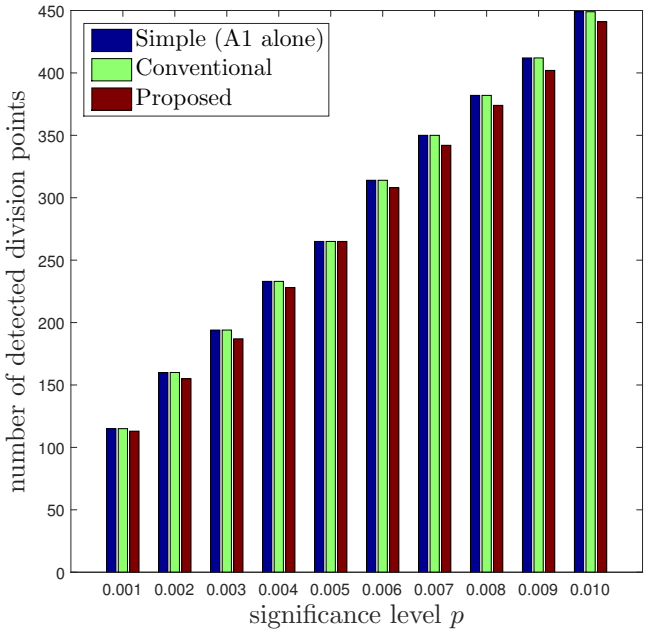

図4.3　各解法による分割点数の結果

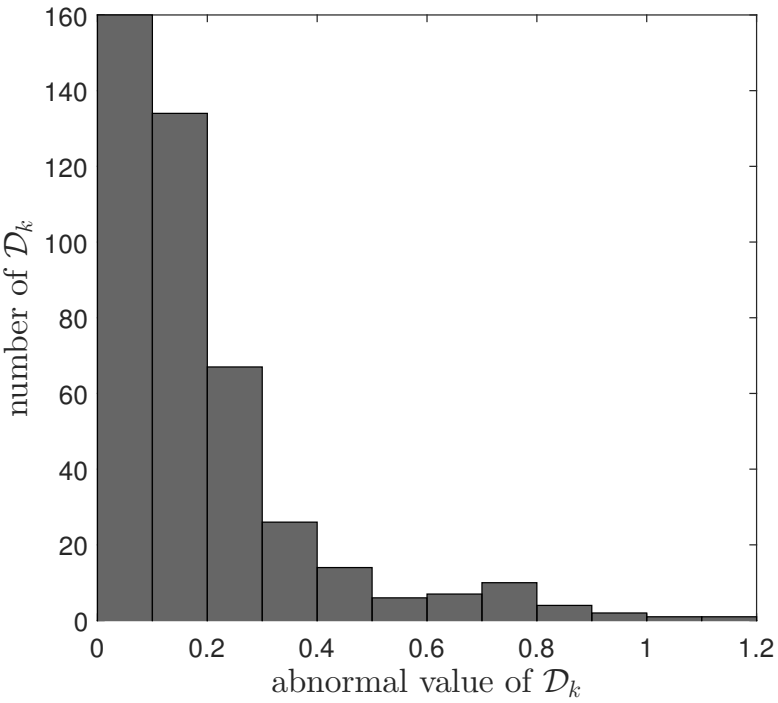

図4.4　異常値の分布

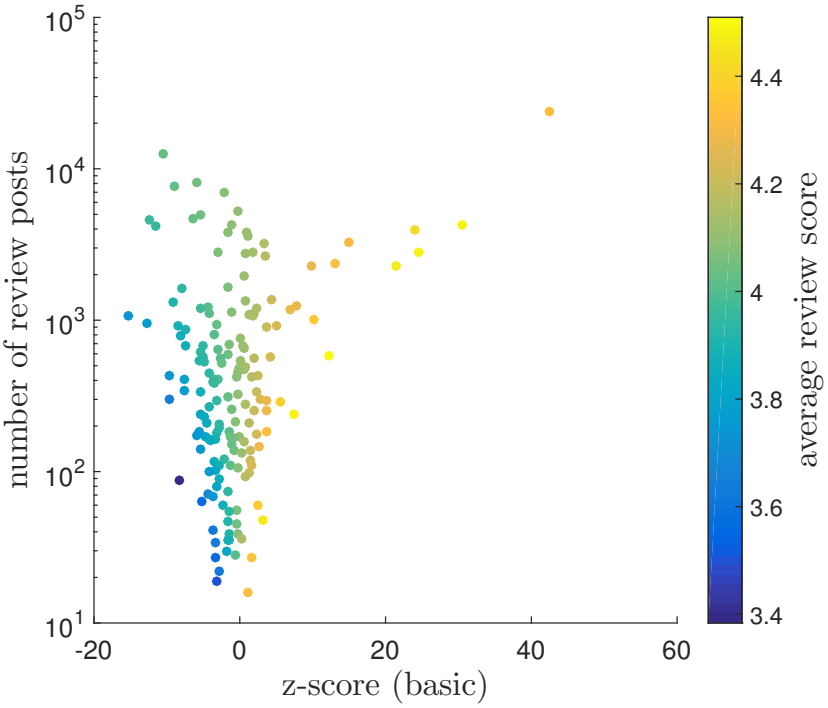

図4.5　区間分割を考慮しない評価値(basic)の分布

4.4.3　評価値の差異

まず，basic' を採用したときの，basic からの評価値の差異の上位を表 4.1に，下位を表 4.2に示す．なお，表に示す $\mathcal{D}_A$ は，異常区間$\mathcal{D}_A = \{\mathcal{D}_k \in \mathcal{D} | \sqrt{|z(\mathcal{D}_k; \hat{\boldsymbol{p}}_k, \mathcal{C}_K)|/|\mathcal{D}_k|} > A\}$ である．代表として，上位で最も分割数が多い「館山」の詳細な結果を図 4.8，4.9，4.10，4.11に示す．図 4.8 は評点データと分割点を示しており，青線は評点の推移，赤線は評点の移動平均の推移，黒い線は分割点をそれぞれ表す．図 4.9 は各区間の評点分布を示しており，横軸は区間 k，縦軸は区間ごとの相対度数，色は評点 j をそれぞれ意味する．図 4.10 は各区間の z-score を示しており，横軸は区間 k，縦軸は z-score $z(\mathcal{D}_k; \hat{\boldsymbol{p}}_k, \mathcal{C}_K)$ をそれぞれ意味する．棒の色の濃淡は，z-score の相対的な位置づけによるものである．図 4.11 は各評価値による結果を示しており，黄色の棒が basic，赤色の棒が basic'，緑色の棒が latest をそれぞれ意味する．「館山」の場合は，最初期の $\mathcal{D}_1$ と，急激に評点分布が変化した短期間の $\mathcal{D}_2$ が異常区間として除外されたことにより，最近の高い評価を示している $\mathcal{D}_4$ に評価が傾いたことが分かる．実際，「館山」は後に示す latest の結果においても，差異の上位に登場する．参考までに，図 4.12 に評価値の差異の分布(basic' - basic)を示す．図より，basic' を採用したときは，レビュー数が比較的少ないアイテムが影響を受けやすいことがわかる．これは，レビュー数が少なければ少ないほど，異常区間

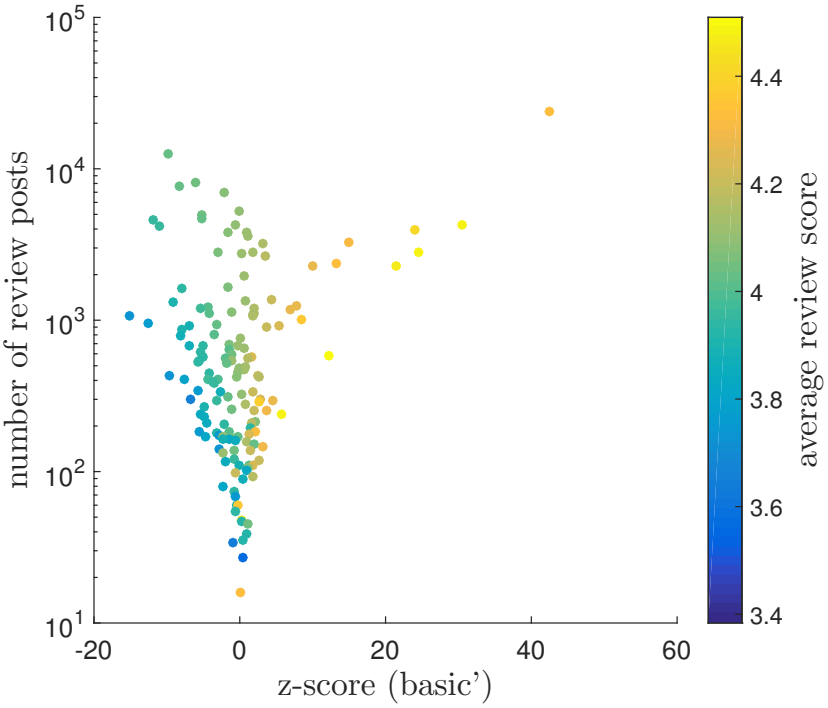

図4.6　異常区間を除いた評価値(basic')の分布

として除去されるレビューの割合が大きくなりやすいことからも容易に推察できる.

表4.1　basic' による評価値の差異の上位

No.	region	N	K	basic	basic'	$\mathcal{D}_A$	diff.
1	小布施	193	2	-2.842	1.477	$\mathcal{D}_1, \mathcal{D}_3$	4.319
2	松前	101	1	-3.387	0.900	$\mathcal{D}_2$	4.287
3	三次	27	1	-3.331	0.422	$\mathcal{D}_1$	3.753
4	松戸	160	1	-4.070	-0.655	$\mathcal{D}_1$	3.415
5	二本松	89	1	-2.785	0.425	$\mathcal{D}_1$	3.210
6	武雄	151	1	-1.155	2.023	$\mathcal{D}_1$	3.178
7	豊橋	173	2	-5.840	-2.874	$\mathcal{D}_2$	2.966
8	伊勢原	68	1	-3.624	-0.662	$\mathcal{D}_2$	2.962
9	柏	163	1	-4.258	-1.369	$\mathcal{D}_1$	2.888
10	館山	303	3	-9.580	-6.699	$\mathcal{D}_1, \mathcal{D}_2$	2.881

次に，latest を採用したときの，basic からの評価値の差異の上位を表 4.3に，下位を表 4.4に示す．代表として，上位で最も分割数が多い「神戸」と「札幌」のうち，図の視認性を考慮して，レビュー数がなるべく少ない「神戸」の詳細な結果を図 4.13，4.14，4.15，4.16に示す．「神戸」の場合は，後半部分に着目すると，評点分布が似ている区間として，$\mathcal{D}_4$ と $\mathcal{D}_6$，$\mathcal{D}_5$ と $\mathcal{D}_7$ の2組に分けることができ

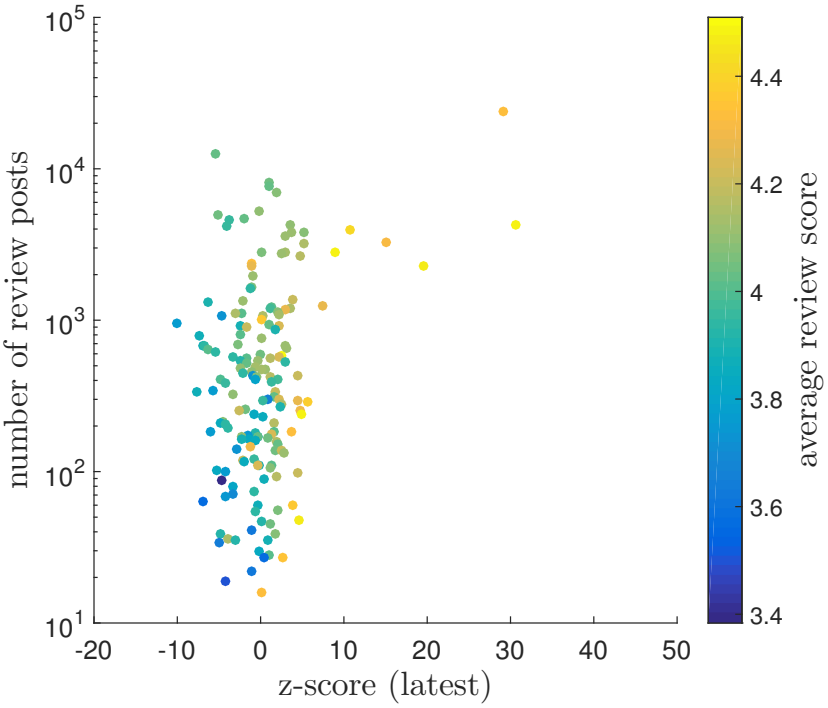

図4.7　最新の時刻区間を用いた評価値(latest)の分布

表4.2　basic' による評価値の差異の下位

No.	region	N	K	basic	basic'	$\mathcal{D}_A$	diff.
1	三鷹	288	1	5.617	2.567	$\mathcal{D}_2$	-3.050
2	新上五島	48	1	3.114	0.265	$\mathcal{D}_2$	-2.849
3	南種子	60	1	2.419	-0.263	$\mathcal{D}_2$	-2.682
4	白川	572	2	4.236	1.702	$\mathcal{D}_2$	-2.534
5	占冠	133	1	0.267	-2.253	$\mathcal{D}_2$	-2.519
6	蔵王	169	1	-0.118	-2.290	$\mathcal{D}_1$	-2.172
7	紋別	99	1	1.249	-0.605	$\mathcal{D}_2$	-1.854
8	渡嘉敷	241	1	7.411	5.643	$\mathcal{D}_2$	-1.768
9	屋久島	1016	2	10.099	8.409	$\mathcal{D}_2$	-1.690
10	東川	182	1	3.735	2.124	$\mathcal{D}_2$	-1.611

る．提案区間分割法が用いる基準は評点分布の変化のみであるが，この場合，評価が比較的悪い時期と良い時期を明確に分けることができていると言える．結果的に，latest を採用することによって，最近の評価の回復を捉えることが可能となっている．参考までに，図 4.17に，評価値の差異の分布(latest - basic)を示す．図より，latest を採用したときは，レビュー数が比較的多いアイテムが影響を受けやすいことがわかる．

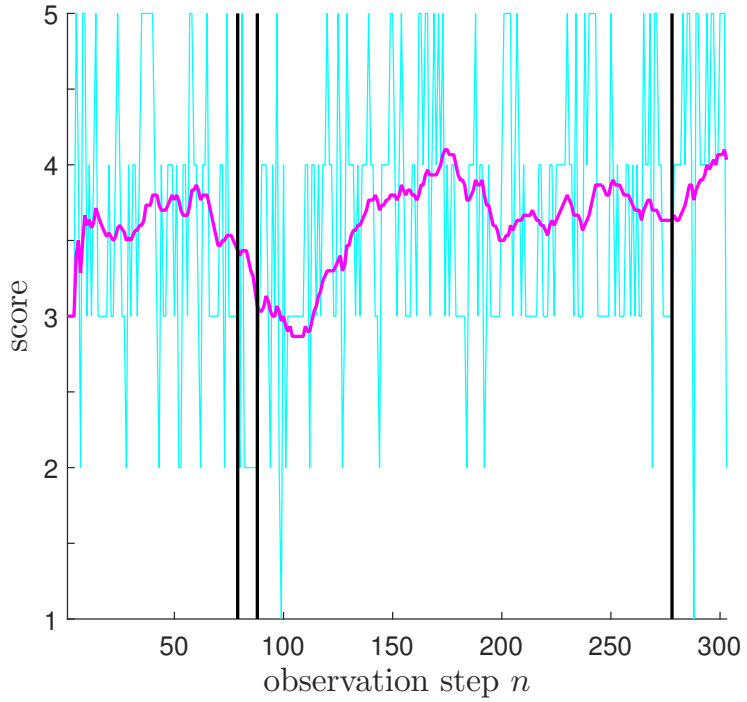

図4.8　館山の評点データと分割点

表4.3　latest による評価値の差異の上位

No.	region	N	K	basic	latest	diff.
1	名護	1064	2	-15.360	-4.731	10.629
2	館山	303	3	-9.580	0.854	10.434
3	札幌	7734	6	-9.029	1.013	10.043
4	那須	865	2	-7.424	1.767	9.190
5	木更津	433	2	-9.584	-0.909	8.675
6	神戸	4630	6	-12.359	-3.729	8.630
7	茅野	532	1	-4.932	3.006	7.937
8	福岡	4204	3	-11.577	-4.077	7.500
9	川越	409	1	-7.619	-0.703	6.915
10	横浜	8122	5	-5.843	1.012	6.855

4.5　おわりに

レビューサイトにおける新たなランキング手法として，提案した区間分割法の結果による区間を z-score 化し，アイテム依存の時期的な信頼を考慮した評価値を提案した．実験において，提案した basic’ (異常区間除去法)と latest (信頼区間採用法)は，basic (基本多項分布法)と比較して，それぞれ

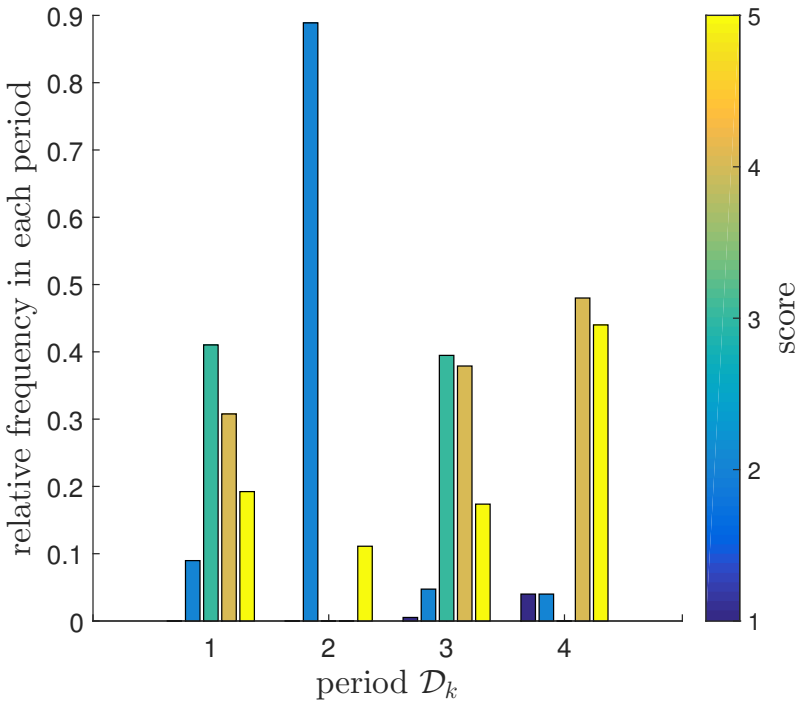

図4.9　館山の各区間の評点分布

表4.4　latest による評価値の差異の下位

No.	region	N	K	basic	latest	diff.
1	本部	2818	2	24.517	8.918	-15.599
2	宮古島	2370	3	13.151	-1.131	-14.282
3	京都	24100	8	42.512	29.164	-13.348
4	浦安	3975	3	24.000	10.727	-13.273
5	伊勢	2277	1	9.862	-1.140	-11.002
6	屋久島	1016	2	10.099	0.160	-9.939
7	上高地	586	1	12.229	2.503	-9.726
8	草津	901	2	3.633	-1.696	-5.329
9	成田	1104	2	1.979	-2.990	-4.970
10	朝来	255	1	1.946	-2.525	-4.471

異なる性質を持つことを示した．特に latest は，レビュー投稿数やレビュー平均評点といったナイーブな情報に依存しにくい性質を持っているため，新たな評価値としての有用性が期待できる．

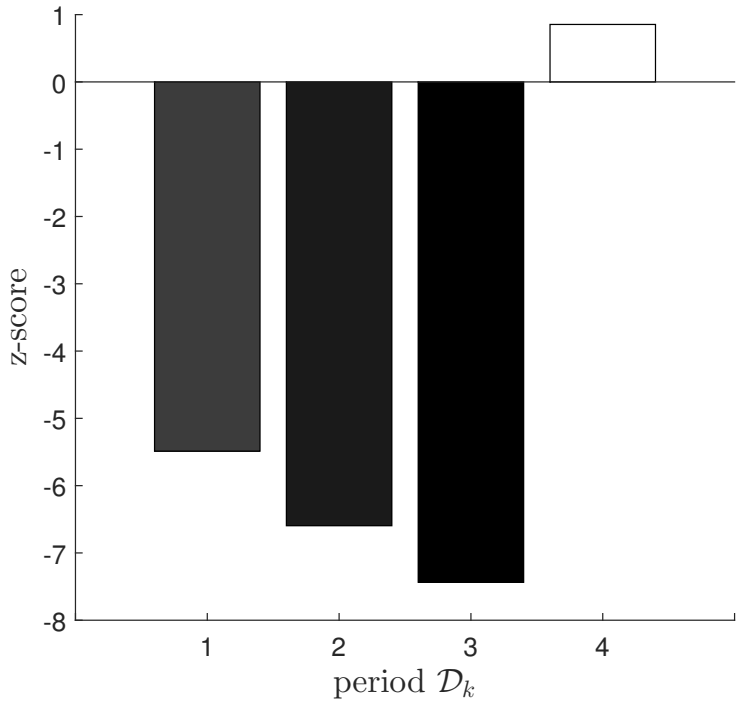

図4.10　館山の各区間の z-score

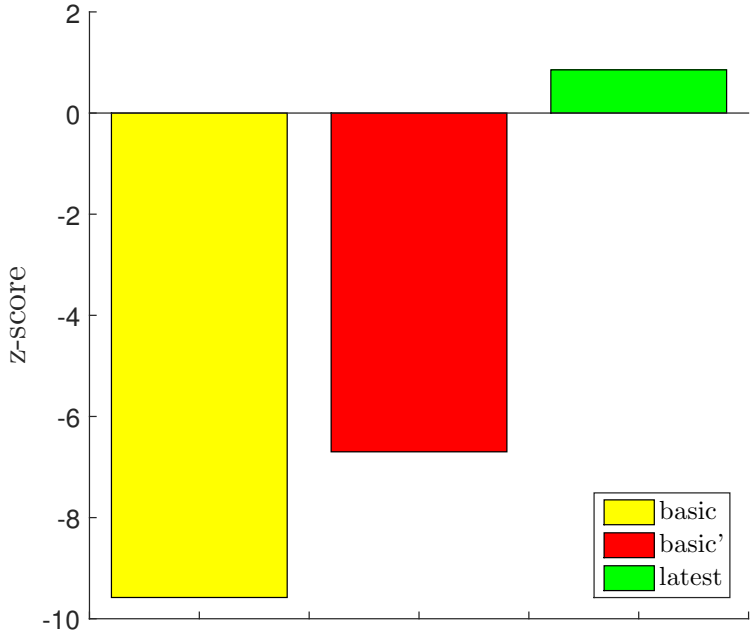

図4.11　館山の評価値結果

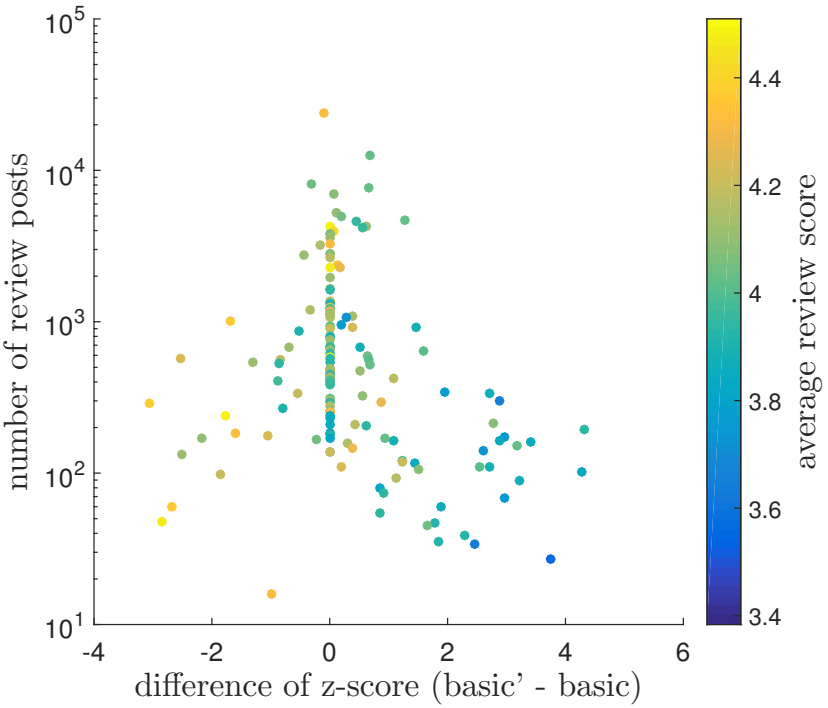

図4.12　評価値の差異の分布(basic' - basic)

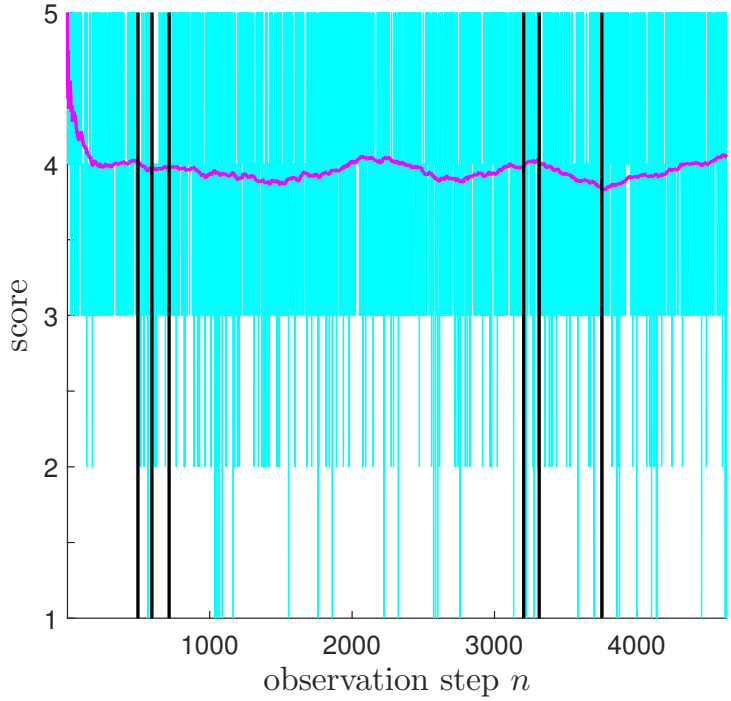

図4.13　神戸の評点データと分割点

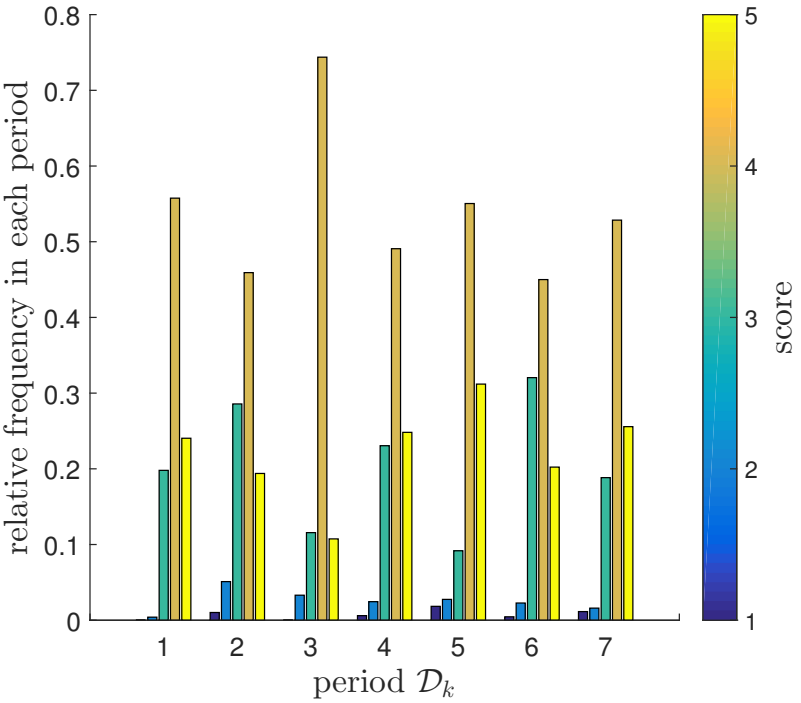

図4.14　神戸の各区間の評点分布

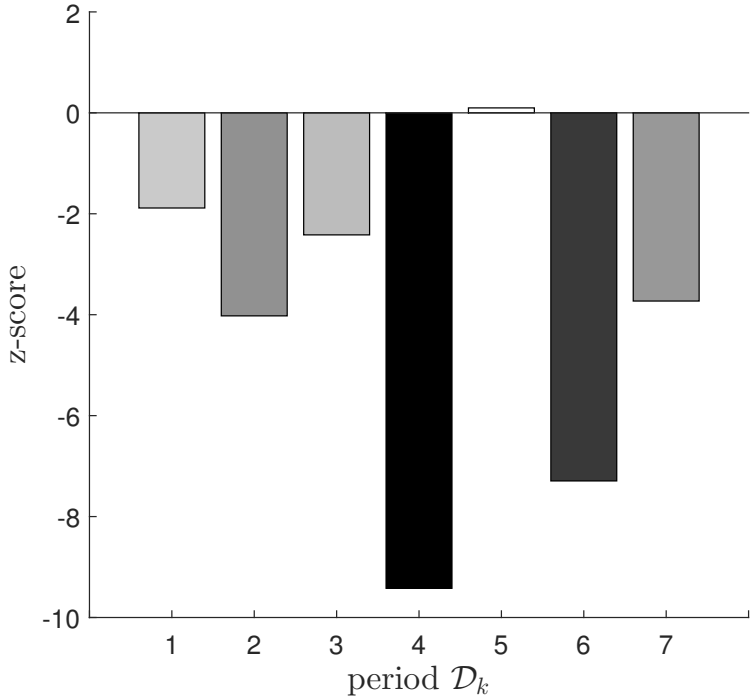

図4.15　神戸の各区間の z-score

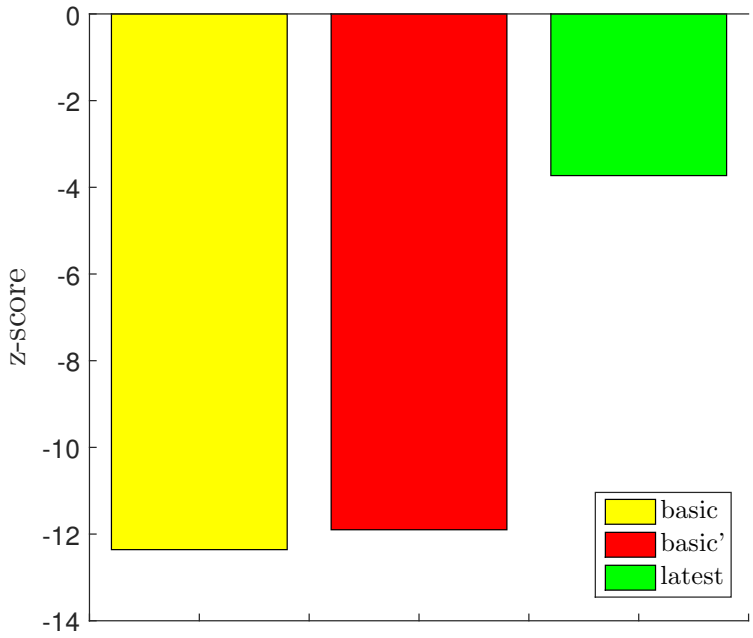

図4.16　神戸の評価値結果

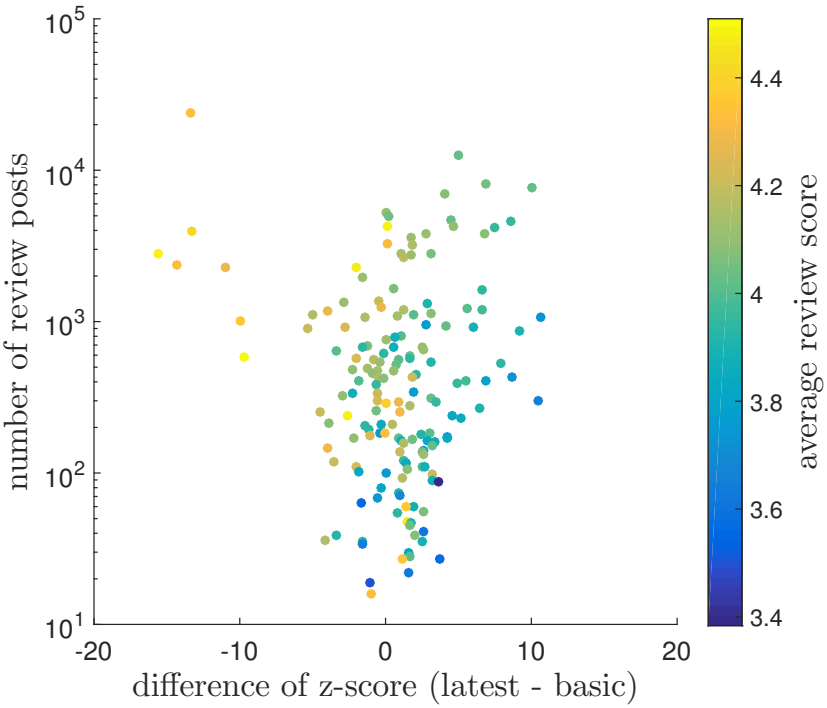

図4.17　評価値の差異の分布(latest - basic)

第5章

カテゴリ評価法：時空間モデル

この章では，位置情報と時刻情報が正確であるデータを扱うことを前提として，第 4章で提案した z-score を時空間モデルとして拡張する．区間分割法を利用したアイテム評価は，各アイテム依存の問題を露呈させることには向いているが，各アイテムの z-score もそれら個々の問題に応じて急激に変化してしまうため，ここでは一度区間分割をしない状態を考える．ただし，区間分割法のときと同様，データの時期的な信頼性を考慮して，z-score に時間的信頼減衰関数を導入する．さらに，位置情報を利用した空間的信頼減衰関数も導入し，順位統計量に基づいたカテゴリ評価の観点から z-score の平等性を向上させることを試みる．

なお，カテゴリ評価法を扱う第 5章から第 7章までは，数式や解法において，章ごとの定義を使用している．

5.1　はじめに

第 4章から続く問題提起となるが，本来ランキングというものは，オブジェクト集合から効率的に高品質なものを見分けるために必要とされている．しかし，レビューサイトでのランキングは，ユーザから提供される情報のみに基づいているため，オブジェクトが登録された時期や，オブジェクトの実際の位置によって，それらの市場における有利不利が生じている可能性が高いということが推察できる．新しいオブジェクトと古いオブジェクトを平等に評価する問題に対しては，時間減衰関数 [13, 14] が頻繁に用いられる．実際，時間減衰の考え方は，ソーシャルメディアマイニングの様々な状況において，既にパフォーマンス向上の功績を収めている．例えば，Koren [15]は，推薦システムにおいて，時間減衰関数を用いたモデルを提案している．加えて，情報拡散過程の時間減衰による影響は，情報拡散モデル上の情報伝播確率の導入において頻繁に扱われている [25, 26, 27]．また，投票者モデル [28, 29]の意見形成モデルにおいても，時間減衰関数を組み込んだ手法が提案されている [30]．当然，今回扱うような観光スポットのレビューデータも，情報の信頼性が登録時期に依存している可能性が高いため，時間減衰関数の有効性が期待できる．更に，今回のデータは正確な位置情報も有しているため，位置による不

平等性問題を考えることもできる．よって本章では，情報の信頼性を考慮することを目的として，時間減衰関数と，それと同様の考え方に基づく空間減衰関数の両方を導入したモデルを構築する．第 4章同様，本章における信頼とはオブジェクト依存のものであるため，推薦システムの研究 [23, 24]で頻繁に用いられているような，ユーザ依存の信頼とは異なることに注意されたい．

5.2　ランキング手法

時刻区間 $\mathcal{T}$ において，整数の評点 $\mathcal{K} = \{1, \cdots, K\}$ によってユーザに評価されたレビュー対象オブジェクトを $\mathcal{V}$ とすると，レビュー集合は $\mathcal{D} = \{(v, k, t) \mid v \in \mathcal{V}, k \in \mathcal{K}, t \in \mathcal{T}\}$ のように書き表せる．任意の $v \in \mathcal{V}$ と $t \in \mathcal{T}$ に対し，時刻 t 以前の時刻 τ からなる v のレビュー集合を $M(v,t) = \{\tau \mid (v,k,t) \in \mathcal{D}, \tau < t\}$ とする．そして，時刻 t におけるオブジェクト v の評点を $g(v,t) \in \mathcal{K}$ とし，$k \in \mathcal{K}$ に対する $M(v,t)$ の部分集合を $M_k(v,t) = \{\tau \in M(v,t) \mid g(v,\tau) = k\}$ とする．いま，過去に投稿された全ての評点を考慮した多項分布モデルを定義する．すなわち，観測されたデータから時刻 t におけるオブジェクト v のレビュー評点分布を予測する以下のモデルを考える．

$$P(g(v,t) = k) = \frac{1 + |M_k(v,t)|}{K + |M(v,t)|}, \ (k = 1, \cdots, K). \tag{5.1}$$

ここで，本手法は Laplace スムージングとして知られるベイズ事前分布を用いた．式 (5.1) の Laplace スムージングは，各オブジェクトが最初に等確率で $1, \cdots, K$ の各評点で評価されたことを仮定している．また，この Laplace スムージングは，ベイズ統計における事前分布として頻繁に用いられるディリクレ分布の特殊ケースに相当しており，実際，ディリクレ分布は多項分布の共役事前分布である．このモデルを基本多項分布モデルとする．

ここから，上記のモデルに基づくオブジェクトランキング手法を提案する．時刻区間 $\mathcal{T}$ における平均評点と標準偏差は，それぞれ $\mu = \sum_{k \in \mathcal{K}} p(k)k$, $\sigma = \sqrt{\sum_{k \in \mathcal{K}} (k-\mu)^2 p(k)}$ のように算出される．ここで，$p(k) = \sum_{v \in \mathcal{V}} |M_k(v,T)| / \sum_{v \in \mathcal{V}} |M(v,T)|$ であり，T は $T = \max\{t \in \mathcal{T}\}$ で定義される最終観測時刻である．各レビュー評点が，評点分布 $p(k)$ に従って独立に与えられたと仮定すると，$\mathcal{Q}$ 個のレビュー $S = \{k_1, \cdots, k_{\mathcal{Q}}\}$ が投稿されたときの期待される平均評点の偏差は，式 (4.3) と式 (4.4) のときと同様の変換で，

$$\begin{aligned} RMSE &= \sqrt{\sum_{k_1 \in \mathcal{K}} \cdots \sum_{k_{\mathcal{Q}} \in \mathcal{K}} \left(\mu - \frac{1}{\mathcal{Q}} \sum_{q=1}^{\mathcal{Q}} k_q \right)^2 \prod_{q=1}^{\mathcal{Q}} p(k_q)} \\ &= \frac{\sigma}{\sqrt{\mathcal{Q}}}, \end{aligned} \tag{5.2}$$

となる．よって，時刻 t におけるオブジェクト v の平均評点の z-score $z(v,t)$ は，以下のように考える

ことができる．

$$z(v,t) = \frac{\mu(v,t) - \mu}{\sigma / \sqrt{|M(v,t)|}},\ \mu(v,t) = \sum_{k \in \mathcal{K}} k \frac{|M_k(v,t)|}{|M(v,t)|}. \tag{5.3}$$

明らかに，この $z(v,t)$ が大きいオブジェクトは，統計的有意に高い評価を得ているとみなすことができる．

ここまで，過去に投稿された全てのレビューは同じ重みであると仮定してきたが，過去と現在で評価の揺らぎがある場合は，古いレビューの信頼度は低くなると考えることができる．更に，位置情報を有するレビュー対象オブジェクトを評価する場合，単純に集合全体の情報を考慮した基準を使用するより，位置が近いオブジェクトの情報を強く，位置が遠いオブジェクトの情報を弱く考慮した基準を使用した方が，地理的な有利不利が起こりにくいことも自然と想定できる．これらの考え方をモデルに反映するために，時空間的信頼減衰関数を導入する．単純な手法としては，$\exp(-\lambda\Delta\cdot)$ のような指数減衰関数が挙げられる．ここで，$\lambda \geq 0$ はパラメータであり，$\Delta\cdot$ は時空間的差異を意味する．

時間的信頼減衰は，オブジェクト v 単体における問題であるとし，その減衰を $\rho_\alpha(\Delta t; \lambda_v) = \exp(-\lambda_v \Delta t)$ で定めるとすると，基本多項分布モデルの式 (5.1) は，

$$P(g(v,t) = k) = \frac{1 + \sum_{\tau \in M_k(v,t)} \rho_\alpha(t - \tau; \lambda_v)}{K + \sum_{\tau \in M(v,t)} \rho_\alpha(t - \tau; \lambda_v)}, \tag{5.4}$$

のように拡張することができる．次に，空間的信頼減衰は，オブジェクト全体 $\mathcal{V}$ における問題であると考える．一般に， インターネット上で得られる位置情報は緯度と経度であるため，GRS80 [31] 準拠楕円体に基づいた，オブジェクト v，w 間の標高を無視した地表面距離を $\Delta d_{v,w}$ とすると，その減衰は $\rho_\beta(\Delta d_{v,w}; \lambda_d) = \exp(-\lambda_d \Delta d_{v,w})$ のように定めることができる．これらの時空間的信頼減衰を考慮し，推定パラメータを $\hat{\lambda}_v$, $\hat{\lambda}_d$ とすれば，各オブジェクト v に対する新たな基準となる評点確率分布は

$$p(v,k) = \frac{\sum_{w \in \mathcal{V}} \sum_{\tau \in M_k(w,T)} \rho_\alpha(t - \tau; \hat{\lambda}_w) \rho_\beta(\Delta d_{v,w}; \hat{\lambda}_d)}{\sum_{w \in \mathcal{V}} \sum_{\tau \in M(w,T)} \rho_\alpha(t - \tau; \hat{\lambda}_w) \rho_\beta(\Delta d_{v,w}; \hat{\lambda}_d)}, \tag{5.5}$$

となり，それに伴い各オブジェクト v に対して期待される平均評点と標準偏差は，それぞれ $\mu(v) = \sum_{k \in \mathcal{K}} p(v,k) k$, $\sigma(v) = \sqrt{\sum_{k \in \mathcal{K}} (k - \mu(v))^2 p(v,k)}$ となる．よって，このときの時刻 t におけるオブジェクト v の平均評点の z-score $z_\rho(v,t)$ は，以下のように置き換えられる．

$$\begin{aligned} z_\rho(v,t) &= \frac{\mu_\rho(v,t) - \mu(v)}{\sigma(v) / \sqrt{\sum_{\tau \in M(v,t)} \rho_\alpha(t - \tau; \hat{\lambda}_v)}}, \\ \mu_\rho(v,t) &= \sum_{k \in \mathcal{K}} k \frac{\sum_{\tau \in M_k(v,t)} \rho_\alpha(t - \tau; \hat{\lambda}_v)}{\sum_{\tau \in M(v,t)} \rho_\alpha(t - \tau; \hat{\lambda}_v)}. \end{aligned} \tag{5.6}$$

この拡張モデルにおける時間的信頼減衰の推定パラメータ $\hat{\lambda}_v$ は，観測データ $\mathcal{D}$ に対するオブジェク

ト v の対数尤度関数，

$$\mathcal{L}(M(v,T);\lambda_v) = \log\left(\prod_{(v,k,t)\in M(v,T)} P(g(v,t)=k)\right), \tag{5.7}$$

を最大化することによって得ることが可能である．この対数尤度関数は，式 (5.4)と $\rho_\alpha(\Delta t;\lambda_v) = \exp(-\lambda_v \Delta t)$ より，

$$\begin{aligned}\mathcal{L}(M(v,T);\lambda_v) = &\sum_{(v,k,t)\in M(v,T)} \log\left(1+\sum_{\tau\in M_k(v,t)} \exp\left(-\lambda_v(t-\tau)\right)\right)\\ &- \sum_{(v,k,t)\in M(v,T)} \log\left(K+\sum_{\tau\in M(v,t)} \exp\left(-\lambda_v(t-\tau)\right)\right),\end{aligned} \tag{5.8}$$

と書き直せる．この尤度関数を最大化するパラメータ $\hat{\lambda}_v$ を EM アルゴリズムで推定する．このとき，Q 関数のヘス行列は半負定値となるため，ニュートン法を用いて Q 関数の大域最適解を得ることができる．また，空間的信頼減衰の推定パラメータ $\hat{\lambda}_d$ は，$\exp(-\lambda_d \Delta d_{v,w})$ の最尤推定量 $1/E(\Delta d_{v,w})$ を用いる．

5.3　実験

5.3.1　データセット

今回使用するデータセットは，TripAdvisor [*1]に登録されている，日本の観光スポットのレビューデータである．このデータセットは，緯度と経度を有する観光スポットのみを扱っており，日本人ユーザのレビュー情報と，英語圏ユーザのレビュー情報で構成されている．取得時期は2014年10月，スポット数 $|\mathcal{V}|$ は11353，総レビュー数 $|\mathcal{D}|$ は296221，レビュー評点は1から5の整数値($k \in \mathcal{K} = \{1,\cdots,5\}$)となっている．参考までに，レビュー評点の相対度数分布を図 5.1に示す．

5.3.2　実験結果

まず，今回のデータセットにおける，各スポット v の時間的信頼減衰関数の推定パラメータ $\hat{\lambda}_v$ と被レビュー数 $|M(v,T)|$ の関係を図 5.2に示す．図から，被レビュー数が多いスポットほど，推定パラメータが低くなる傾向が見て取れる．つまり，被レビュー数が多いにも関わらず推定パラメータが高いスポットは，直近と過去で評価の揺らぎが激しいことが予想できる．次に，今回のデータセットにおけるオブジェクト間距離 $\Delta d_{v,w}$ の相対度数分布を図 5.3に示す．そして，データセットから求めた最尤推定量 $\hat{\lambda}_d$ を用いた空間的信頼減衰関数 $\exp(-\lambda_d \Delta d_{v,w})$ を示したのが図 5.4である．今回の拡張モデル

[*1] http://www.tripadvisor.com/

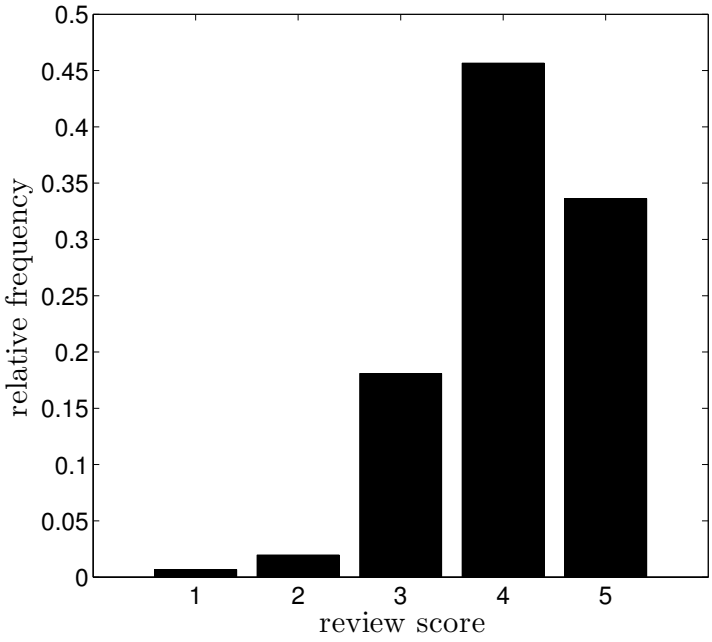

図5.1　レビュー評点の相対度数分布

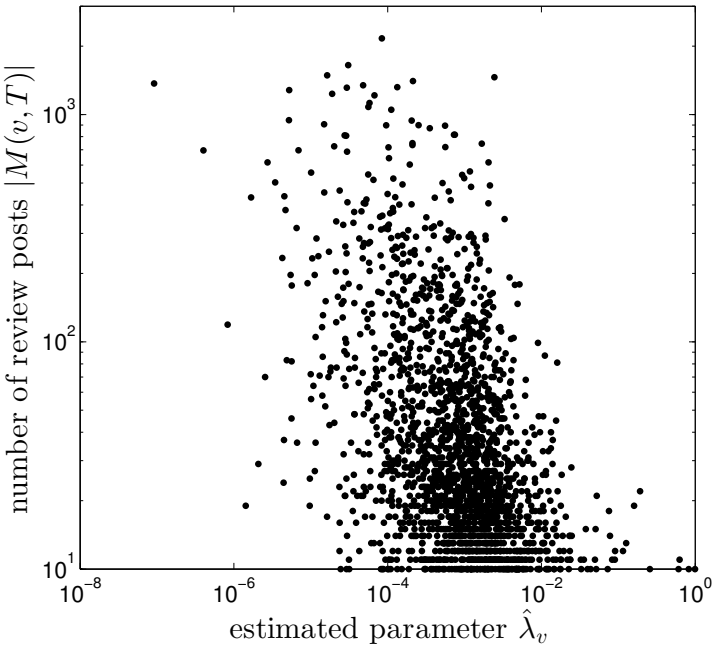

図5.2　推定パラメータ $\hat{\lambda}_v$ と被レビュー数 $|M(v,T)|$ の関係

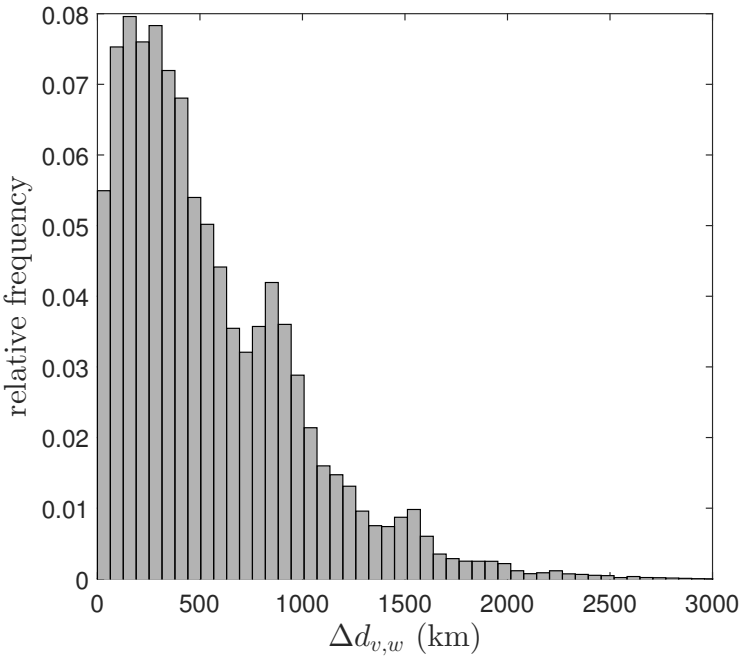

図5.3　オブジェクト間距離 $\Delta d_{d,w}$ の相対度数分布

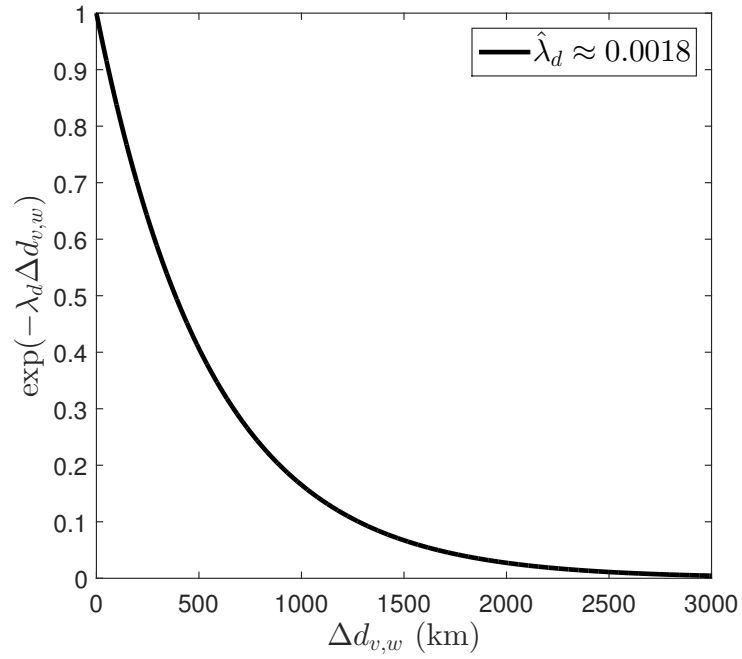

図5.4　最尤推定量 $\hat{\lambda}_d$ と空間的信頼減衰関数 $\exp(-\lambda_d \Delta d_{v,w})$

に基づき，これらの推定パラメータと減衰関数を用いて算出した期待平均評点 $\mu(v)$ の分布は図 5.5のようになった．

続いて，基本多項分布モデルに基づいて算出した各スポットの評価値 $z(v,T)$ の分布を図 5.6に，拡張モデルと推定パラメータに基づいて算出した各スポット v の評価値 $z_\rho(v,T)$ の分布を図 5.7にそれぞれ示す．図より，両評価値共に，投稿されたレビュー数 $|M(v,T)|$ が多くなるほど評価値の幅が広がるようになっており，単にレビュー平均評点 $\mu(v,T)$ が高い（又は低い）だけで評価値が極端に高く（又は低く）なっていないことがわかる．特に，$z_\rho(v,T)$ は，過去に投稿されたレビューの影響度と，評価の基準となる期待平均評点 $\mu_\rho(v)$ が v 毎に変化するため，レビュー投稿数が同程度でも，評価値の大小がレビュー平均評点 $\mu(v,T)$ に完全に準じていないことに注意されたい． 本章では，基本多項分

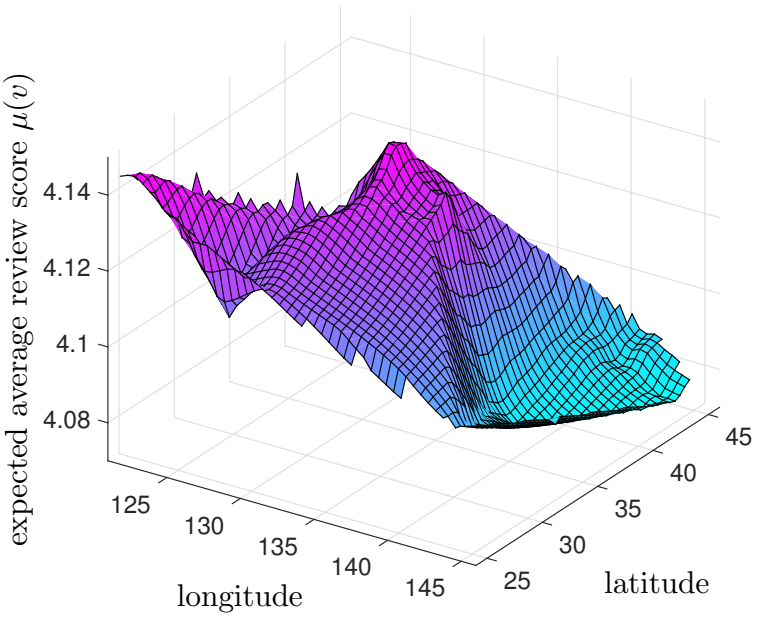

図5.5　推定パラメータと拡張モデルに基づく期待平均評点 $\mu(v)$ の分布

布モデルに基づく $z(v,T)$ によるランキングを単純法 (*simple*)，拡張モデルと推定パラメータに基づく $z_\rho(v,T)$ によるランキングを提案法 (*proposed*) とし，それぞれのランキングの地理的な平等性を定量的に評価する．この評価には，以下に述べるカテゴリ評価法の評価値の分散を用いる．

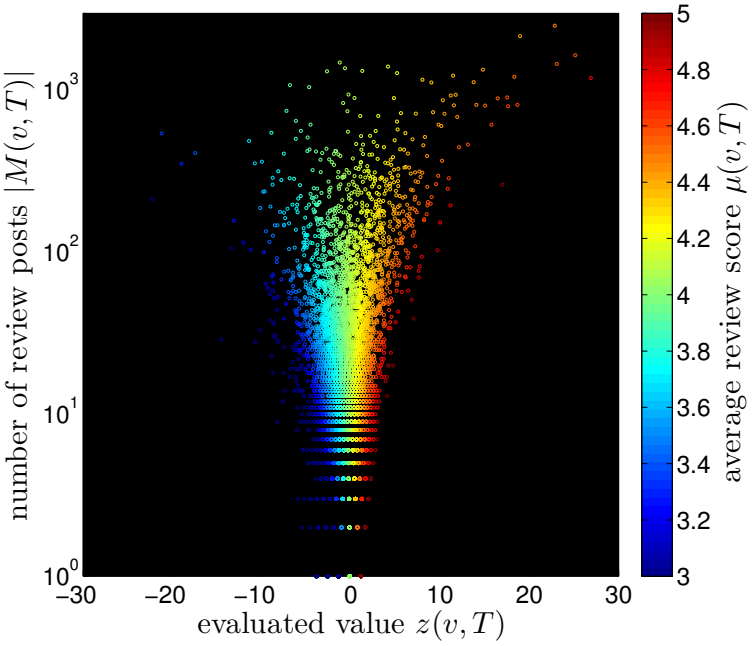

図5.6　投稿されたレビュー数 $|M(v,T)|$ とレビュー平均評点 $\mu(v,T)$ と基本評価値 $z(v,T)$ の関係

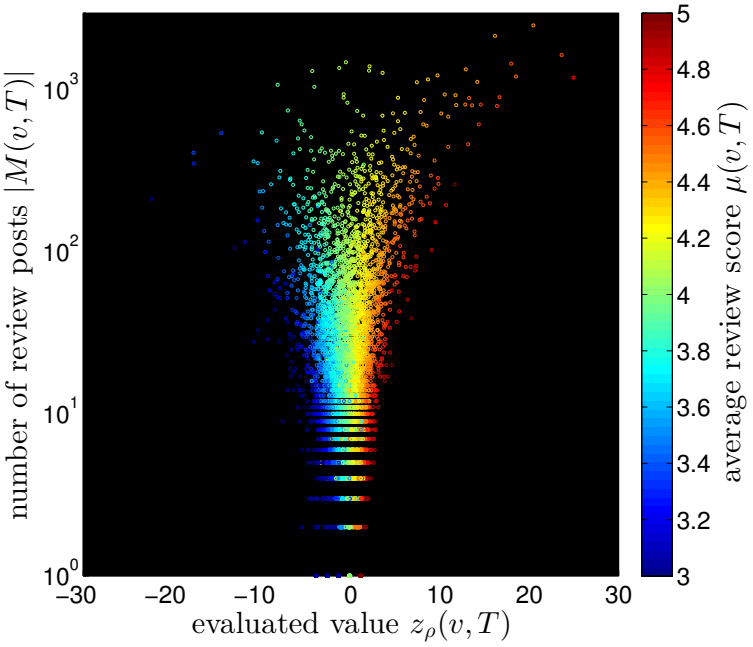

図5.7　投稿されたレビュー数 $|M(v,T)|$ とレビュー平均評点 $\mu(v,T)$ と提案評価値 $z_\rho(v,T)$ の関係

5.4　カテゴリ評価法

5.4.1　問題設定

与えられたオブジェクト集合とカテゴリ集合をそれぞれ $\mathcal{I}$ と $\mathcal{J}$ とする．ここで，それぞれの要素数は $I = |\mathcal{I}|$ と $J = |\mathcal{J}|$ とし，各要素は整数と同一視されるとする．つまり，$\mathcal{I} = \{1, \cdots, i, \cdots, I\}$ および $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$ とする．また，オブジェクト i が属すカテゴリを $j = f(i)$ で表し，各カテゴリに属すオブジェクト数を $I_j = |\mathcal{I}_j| = |\{i\ ;\ j = f(i)\}|$ とする．各オブジェクト i に対し，そのランキングは $1 \leq r_i \leq I$ で与えられるとする．ただし，同順位が起こるケースでは，r_i は平均順位で補正されるとする．

ここでの目的は，カテゴリとランキング付きのオブジェクトの集合が与えられたとき，ランキングの高い，または逆に低いオブジェクトが有意に多く含まれるカテゴリを定量的に評価する指標の構築である．以下には，Mann-Whitney の統計量 [32] に基づく自然な拡張法を示す．

5.4.2　多群順位統計量

Mann-Whitney の二群順位統計量を多群に拡張して適用する方法について述べる．いま，カテゴリ j に着目すれば，このカテゴリに属すオブジェクト集合 $\mathcal{I}_j$ と，それ以外のオブジェクト集合 $\mathcal{I} \setminus \mathcal{I}_j$ の二群に分割することができる．ここで，$\cdot \setminus \cdot$ は集合差を意味する．よって，Mann-Whitney の二群順位統計量に従い，次式により，カテゴリ j に対し z-score $\dot{z}_j$ を求めることができる．

$$\dot{z}_j = \frac{\dot{u}_j - \dot{\mu}_j}{\dot{\sigma}_j}. \tag{5.9}$$

ここで，統計量 $\dot{u}_j$, 順位の平均 $\dot{\mu}_j$，および，その分散 $\dot{\sigma}_j^2$ は次のように計算される．

$$\dot{u}_j = I_j(I - I_j) + \frac{I_j(I_j + 1)}{2} - \sum_{i \in \mathcal{I}_j} r_i, \tag{5.10}$$

$$\dot{\mu}_j = \frac{I_j(I - I_j)}{2}, \tag{5.11}$$

$$\dot{\sigma}_j^2 = \frac{I_j(I - I_j)(I + 1)}{12}. \tag{5.12}$$

ただし，同順位が起こるケースでは，標準偏差 $\dot{\sigma}_j$ は標準的な方法で補正されるとする．よって，式 (5.9) で求まる z-score $\dot{z}_j$ により，各カテゴリ j がランキングの高い，または逆に低いオブジェクトを有意に多く含むか定量的に評価することができる．

既に述べているように，この多群順位統計量は，基本的には2クラス分類器の SVM (Support Vector Machine) [33] を多クラス分類器に拡張するときに利用される one-against-all と類似した考え方となる．

5.4.3　ランキング比較

各スポットを i，TripAdvisor において定められている地域区分をカテゴリ j としたときの，単純法，提案法のそれぞれのランキングにおけるカテゴリ評価値 $\dot{z}_j$ の分布を図 5.8, 5.9に示す．この分散が大きい（又は小さい）ということは，ランキングの上位と下位で地域差が大きい（又は小さい）と考えることができる．両図より，カテゴリ評価値の分散は，単純法において 1.3177，提案法において 1.2564 であるため，提案法の方がカテゴリ評価値の散らばりを抑えられていることがわかる．つまり，提案評価値は，基本評価値と比べて地理的な不平等性が低いと言える．この差についての比較を行うため，提案法において時間的信頼減衰を考慮しない，即ち $\hat{\lambda}_v = 0$ で固定した *spatial* と，提案法において空間的信頼減衰を考慮しない，即ち $\hat{\lambda}_d = 0$ で固定した *temporal* で同様の実験を行った．これら4手法におけるカテゴリ評価値 $\dot{z}_j$ の分散比較を示したのが図 5.10である．図より，spatial と temporal も，単純法と比較して地域差が小さくなっているが，これら4手法の比較においても提案法が最も優れていることがわかる．

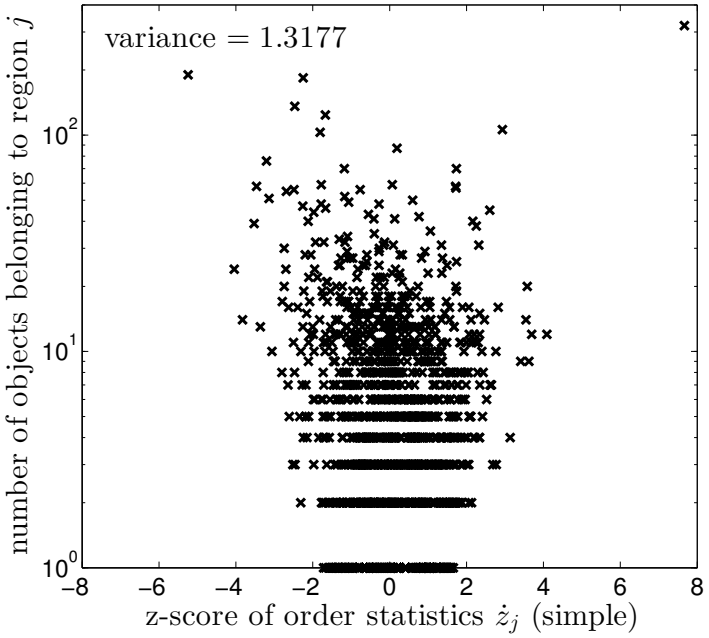

図5.8　単純法におけるカテゴリ評価値 $\dot{z}_j$ の分布

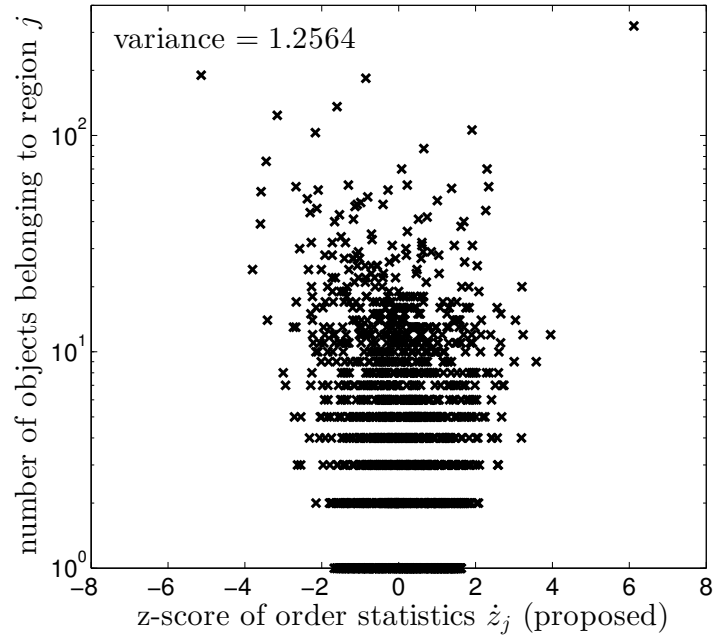

図5.9　提案法におけるカテゴリ評価値 $\dot{z}_j$ の分布

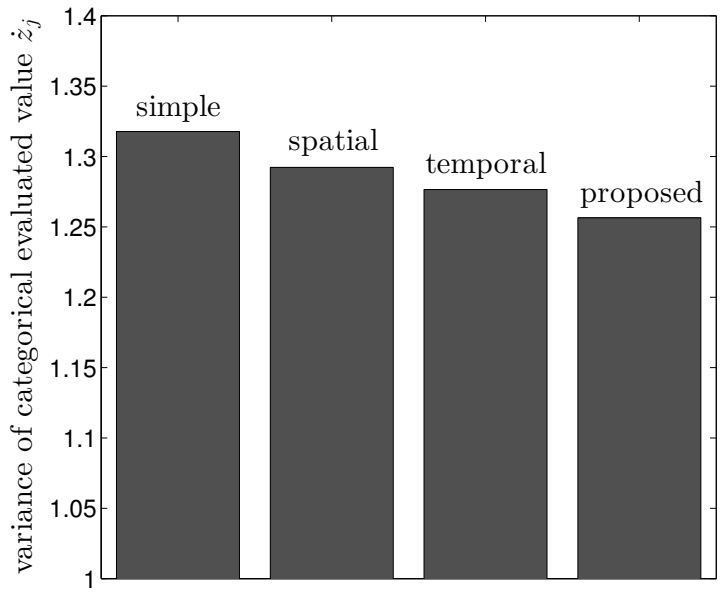

図5.10　カテゴリ評価値 $\dot{z}_j$ の分散比較

5.5　おわりに

位置情報と時刻情報が正確であるデータを扱うことを前提として，第 4章で提案した z-score に対し，情報の時間的信頼を考慮することを目的として時間減衰関数を，情報の地理的信頼性を考慮することを目的として空間減衰関数を導入した．ランキング結果の分析においては，二群順位統計量を多群に拡張し，各カテゴリを評価する方法を提案した．時空間的信頼減衰を考慮したモデルに基づくアイテムランキングは，単純な多項分布モデルに基づくアイテムランキングと比較して，地域カテゴリによる不平等性が低いことを示した．

第6章

カテゴリ評価法：ネットワークモデル

本章では，第 5章で述べた時空間モデルをさらに発展させたネットワークモデルについて述べる．このネットワークモデルは，各データオブジェクトを完全ネットワークのノードとして扱い，時空間情報をネットワーク内の移動情報として利用するものである．ただし，その移動情報はノード間の距離とノードの人気度（知名度・認知度）に依存していると仮定し，距離に関するパラメータと人気度に関するパラメータは機械学習的手法で推定する．ここでは，このネットワークモデルにおけるノードの評価指標と，それらを使ったカテゴリ評価方法について提案する．

なお，数式や解法においては，本章で全て定義し直すものとする．

6.1 はじめに

TripAdvisor [*1] に代表されるような，観光に関するソーシャルメディアの出現によって，近年，あらゆる観光スポットに対してレビューが投稿されるようになった．インターネット上のレビューは，ユーザの個々の意思決定に基いて常に生成され続けているため，大規模なレビューデータに対しては，人間行動の本質的な特性，すなわちいくつかの統計的な規則性が自然と仮定できる．よって，巨視的分析の見地から，人間の観光行動の統計的規則性を見出すことは可能なはずであり，それらの統計的性質に基いて確率モデルを構築すれば，人間の観光行動における精緻な予測が期待できる．特に，そのような行動予測を利用した地域分析は，社会動向や市場動向の調査のためにも重要であると言える．従って本章では，人間の行動特性を利用した観光レビューデータの分析手法を，カテゴリ評価の観点から提案する．

本章では，ユーザ行動のモデル化の先行研究として，Lévy flight [34, 35, 36, 37] に焦点を当てる．実際，今回提案する確率モデルは Lévy flight の特性を大いに利用しており，観光スポットの人気度もこの特性に関連付けて扱っている．よって，各観光スポットの人気度を導入した Lévy flight に基いた確率モデルと，観測されたユーザ行動データから，そのモデルのパラメータを推定する効率的な学習アルゴリズムを提案する．提案手法は，観測されたユーザの観光行動について，ユーザが次にどこの観光

[*1] http://www.tripadvisor.com/

スポットへ移動するかという予測を最適化させることにより，このモデルのパラメータを推定するものである．より具体的には，観測されたユーザ行動データについての対数尤度関数を構築し，機械学習の枠組みで関数の最大化を行う [38]．提案学習アルゴリズムは，尤度関数の凸性を十分に利用した反復計算の仕組みを用いているため，効率的に大域最適解を求めることが可能である [39]．更に，このユーザ行動の確率モデルの応用として，スポット間に条件付き確率を持つ有向リンクを張った，観光スポット間の確率ネットワークを構築する．また，ネットワーク上の重要ノードを見分ける研究 [17] と同じように，それらの条件付き確率に基づいた2 種類のランキング手法を提案する．各ランキング結果は，Mann-Whitney の統計量 [32] に基づき，都道府県区分によって評価する．

提案モデルと提案ランキングの評価には，TripAdvisor のデータセットから生成したユーザ行動データを用いる．具体的には，まず，ユーザ行動データの基本統計量を示した後，観光スポットの人気度のスケールフリー性 [40] と行動データにおける移動距離のスケールフリー性 [34, 40] を調べる．そして，パラメータ推定に関する実験結果を述べ，提案したランキング手法とナイーブな人気度ランキングとの比較を行う．本章における提案モデルと提案ランキング手法は，orienteering problem [41] 等に対する他の手法を改善するためのコア技術にもなる得ることが考えられる．

6.2　関連研究

昨今の技術革新と高性能なモバイルデバイスの普及は，現代人のコミュニケーションスタイルを大きく変え，種々のソーシャルメディアは現代人の日常生活に多大な影響を及ぼしている．よって，近年は情報推薦システム [2] の研究が注目されており，本章の研究内容はロケーションベースの推薦手法 [42] に該当すると言える．特に注目すべき研究としては，Zheng らが提案したGPS の軌跡から重要なスポットを発見する手法 [43] が挙げられる．しかし，ほとんどの既存法は，人間の行動特性を仮定することなく考案されているため，予測性能の改善は困難なことが予期される．

人間行動の基本的な特性としては，人間の移動距離の分布は$p(\delta) \propto \delta^{\beta}$ のような冪乗則によって近似できることが報告されている [34, 40]．ここで，δ と β はそれぞれ移動距離と指数パラメータである．これは，人間の移動パターンは，Lévy flight によってモデル化が可能であることを意味している．本章では，観測されたユーザ行動に対して，各観光スポットの人気度を導入した Lévy flight に基いた確率モデルを提案する．Lévy flight において必要となる β のようなモデルパラメータを推定するために，ここでは対数尤度関数に基づく統計的機械学習手法 [38] を採用し，この関数を最大化するために非線形最適化における反復アルゴリズム [44] を使用する．本章におけるモデルは，目的関数の凸性によって大域最適解を持つことが保証されている [39] ことに注意されたい．

提案モデルの推定パラメータが求まれば，精緻な予測性能を備えた観光スポット間の大規模ネットワークを構築することが可能となる．かねてより，大規模な複雑ネットワークの構造や機能に関する研

究は，社会学，生物学，物理学，コンピュータ科学等の様々な分野で注目されている [16]．特に，これらのネットワークにおけるスケールフリー性は幅広く研究されており [40, 45]，次数相関 [46] 等のより複雑な特徴が提案されてきた．本章では，ネットワークが持つとされるこれらの特徴に着目し，データセットと実験結果の分析を行う．

更に今回は，構築した確率ネットワークを用いて，日本の観光における地域毎の重要度を評価することを試みる．一般に，様々な側面から重要ノード群を発見することは，ネットワーク分析において基礎的な問題とされている．ソーシャルネットワーク分析の分野では，次数中心性，近接中心性，媒介中心性といった，中心性の指標が幅広く研究されており [17]，一方，Web 情報検索の分野では，PageRank [18] とHITS [19] によるノードランキングが広く用いられている．これらの手法の中でも，次数に関する指標とPageRank は今回の確率ネットワークに適用することができるため，本章ではこの2 手法に基づいたランキングを提案する．実験では，ナイーブな人気度ランキングと比較して，提案ランキング手法によって地域毎の評価がどのように変化するかを検証する．

6.3　提案手法

まず，ユーザ行動に対する確率モデルを提案する．ユーザ集合を $\mathcal{U} = \{u, v, w, \cdots\}$，スポット集合を $\mathcal{S} = \{q, r, s, \cdots\}$ とし，それぞれの要素数を $N = |\mathcal{U}|$, $M = |\mathcal{S}|$ とする．ここで，2 つのスポット r, s 間の距離を $d(r, s)$ として表す．そして，指数パラメータ θ_1 によるLévy flight に従えば，ユーザ u がスポット r を訪れた後にスポット s を訪れる条件付き確率$p_1(s \mid r; \theta_1)$ は，$d(r, s)^{-\theta_1}$ に比例することが仮定できる．すなわち，その関係は次式となる．

$$p_1(s \mid r; \theta_1) = \frac{d(r, s)^{-\theta_1}}{\sum_{q \in \mathcal{S}} d(r, q)^{-\theta_1}}. \tag{6.1}$$

後のデータセットの分析で明らかにするが，今回のデータのスポット人気度はスケールフリー性を持っている [40] ため，指数パラメータ θ_2 を用いたスポット $s \in \mathcal{S}$ の人気度を $f(s)$ とすれば，ユーザ u がスポット s に訪れる確率 $p_2(s; \theta_2)$ は $f(s)^{\theta_2}$ に比例することが仮定できる．すなわち，その関係は次式となる．

$$p_2(s; \theta_2) = \frac{f(s)^{\theta_2}}{\sum_{q \in \mathcal{S}} f(q)^{\theta_2}}. \tag{6.2}$$

よって，これらの確率 $p_1(s \mid r; \theta_1)$, $p_2(s; \theta_2)$ を組み合わせれば，ユーザの基本行動モデルとして，以下の条件付き確率を得ることができる．

$$\begin{aligned} p(s \mid r; \boldsymbol{\theta}) &= \frac{p_1(s \mid r; \theta_1) p_2(s; \theta_2)}{\sum_{q \in \mathcal{S}} p_1(q \mid r; \theta_1) p_2(q; \theta_2)} \\ &= \frac{d(r, s)^{-\theta_1} f(s)^{\theta_2}}{\sum_{q \in \mathcal{S}} d(r, q)^{-\theta_1} f(q)^{\theta_2}}. \end{aligned} \tag{6.3}$$

$\boldsymbol{\theta}$ は $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ であり，$\boldsymbol{a}^T$ はベクトル $\boldsymbol{a}$ の転置を意味する．ここで，本章のモデルは，他の要因に基づく訪問確率$p(s \mid \theta)$ を導入することによって，容易に拡張が可能であることを強調しておきたい．

次に，パラメータベクトル $\boldsymbol{\theta}$ を推定する学習アルゴリズムについて述べる．ユーザ $u \in \mathcal{U}$ がスポット $s \in \mathcal{S}$ に時刻 t で訪れたことを (u, s, t) で表せば，観測されたユーザ行動データは $\mathcal{D} = \{\cdots, (u, s, t), \cdots\}$ のように書ける．観測データ $\mathcal{D}$ から，ユーザ u が m 番目に訪問したスポットが分かるため，それを $s(u, m) \in \mathcal{S}$ として表す．ここからは，$M(u)$ をユーザ u が訪れたスポット数，$N(s)$ をスポット s に訪れたユーザ数とする．$\mathcal{D}$ についての $\boldsymbol{\theta}$ を推定するために，標準的な機械学習アプローチ [38] に基づいて，最大化する目的関数として以下の対数尤度関数を考える．

$$L(\boldsymbol{\theta}; \mathcal{D}) = \sum_{u \in \mathcal{U}} \sum_{1 \le m < M(u)} \log p(s(u, m+1) \mid s(u, m); \boldsymbol{\theta}). \tag{6.4}$$

そして，$\boldsymbol{x}(r, s) = (-\log d(r, s), \log f(s))^T$ のように定義されたベクトルを新たに導入すれば，式 (6.3) より，式 (6.4) は以下のように変形できる．

$$L(\boldsymbol{\theta}; \mathcal{D}) = \sum_{u \in \mathcal{U}} \sum_{1 \le m < M(u)} \left(\boldsymbol{\theta}^T \boldsymbol{x}(s(u, m), s(u, m+1)) - \log \sum_{q \in \mathcal{S}} \exp(\boldsymbol{\theta}^T \boldsymbol{x}(s(u, m), q)) \right). \tag{6.5}$$

よって，式 (6.4) で定義された目的関数の勾配ベクトルとヘス行列が以下のように計算できる．

$$\frac{\partial L(\boldsymbol{\theta}; \mathcal{D})}{\partial \boldsymbol{\theta}} = \sum_{u \in \mathcal{U}} \sum_{1 \le m < M(u)} \left(\boldsymbol{x}(s(u, m), s(u, m+1)) - \sum_{q \in \mathcal{S}} p(q \mid s(u, m); \boldsymbol{\theta}) \boldsymbol{x}(s(u, m), q)) \right),$$

$$\frac{\partial^2 L(\boldsymbol{\theta}; \mathcal{D})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \sum_{u \in \mathcal{U}} \sum_{1 \le m < M(u)} - \left(\sum_{q \in \mathcal{S}} p(q \mid s(u, m); \boldsymbol{\theta}) \boldsymbol{x}(s(u, m), q) \boldsymbol{x}(s(u, m), q)^T \right.$$
$$\left. - \left(\sum_{q \in \mathcal{S}} p(q \mid s(u, m); \boldsymbol{\theta}) \boldsymbol{x}(s(u, m), q)) \right) \left(\sum_{q \in \mathcal{S}} p(q \mid s(u, m); \boldsymbol{\theta}) \boldsymbol{x}(s(u, m), q)) \right)^T \right).$$

ここで，このヘス行列は穏やかな条件下で負定値となることから，目的関数が上に凸な単峰関数であることが分かるため，本手法のモデルが大域最適解を持つことが保証される [39]．従って，任意の初期パラメータ値から始まるような反復計算を用いることが可能である [44]．実験では，次式の修正ベクトルによるニュートン法を用いる．

$$\boldsymbol{\delta} = -\frac{\partial L(\boldsymbol{\theta}; \mathcal{D})}{\partial \boldsymbol{\theta}} \left(\frac{\partial^2 L(\boldsymbol{\theta}; \mathcal{D})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1}. \tag{6.6}$$

終了条件として $\epsilon = 10^{-8}$ を設定すれば，学習アルゴリズムは以下のように要約できる．

1. パラメータベクトルを $\boldsymbol{\theta}_v \leftarrow \mathbf{0}$ と初期化する．
2. 式 (6.6) で修正ベクトル $\boldsymbol{\delta}$ を計算し，もし $\|\boldsymbol{\delta}\| < \epsilon$ となれば反復を終了する．
3. パラメータベクトルを $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{\delta}$ と更新し，手順 2. に戻る．

最後に，提案行動モデルに基づいた応用について述べる．学習アルゴリズムによる推定パラメータ値を $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathcal{D})$ とすると，スポット r からスポット s に訪れる条件付き確率$p(s \mid r; \hat{\boldsymbol{\theta}})$ を得ることができる．従って，各リンク $(r, s) \in \mathcal{S} \times \mathcal{S}$ に条件付き確率 $p(s \mid r; \hat{\boldsymbol{\theta}})$ を割り当てたスポット間ネットワーク $G = (\mathcal{S}, \mathcal{S} \times \mathcal{S})$ を考えることができる．先に述べたように，複雑ネットワークにおける基礎的問題の一つは，与えられたネットワークに対して有用な指標を使い，重要ノード（今回の場合は観光スポット）を探索することである．よって，ここでは入次数とPageRank を参考にした2 種類の指標を考える．入次数に基づく手法（入次数法）では，各スポット $s \in \mathcal{S}$ の指標を以下のように定義する．

$$id(s) = \sum_{q \in \mathcal{S}} p(s \mid q; \hat{\boldsymbol{\theta}}). \tag{6.7}$$

すなわち，この手法では，他のスポットからの訪問確率の合計値が大きいスポットほど上位となる．ここで，他のスポットへの訪問確率の合計は $\sum_{q \in \mathcal{S}} p(q \mid s; \hat{\boldsymbol{\theta}}) = 1$ になることに注意されたい．一方，PageRank に基づく手法（PageRank 法）では，ランダムウォーク過程下での訪問確率が高いスポットほど上位となる．PageRank アルゴリズム [18] に従えば，スポット $s \in \mathcal{S}$ の訪問確率 $pr(s)$ は以下のように考えられる．

$$pr(s) \leftarrow (1 - \alpha) \sum_{q \in \mathcal{S}} p(s \mid q; \hat{\boldsymbol{\theta}}) pr(q) + \frac{\alpha}{M}. \tag{6.8}$$

ここで，α は一様ジャンプ確率であり，一般的な値として$\alpha = 0.15$ と設定した [18]．上記のランダムウォークシミュレーションを行うことによって，ランキングの指標となる定常状態値 $pr(s)$ を得ることができる．

6.4　データセット

現実データとして, TripAdvisor から日本の観光スポット, 及びそれらに投稿された日本語のレビューを取得し，レビューの順序に基いてユーザ行動データ $\mathcal{D}$ を構築した．まず，このデータセットの基本統計量について述べる． このデータセットは，$441,087$ レビュー，$N = 52,355$ ユーザ，$M = 19,827$ スポットを有しており，対象期間は2007/02/07 から2015/10/21 迄である．よって，ユーザが投稿したレビュー数の平均は 8.4，スポットに投稿されたレビュー数の平均は 22.2 である．また，レビュー評点は整数の 1 から 5 であり，全評点の平均点は 4.0 である．図 6.1 は 2008 年からの1ヶ月毎のレビュー投稿数の推移を示している．数が大きく変動しているが，投稿数は着実に増加していることが図から見て取れる．図 6.2 はレビュー評点の度数分布を示している．図から，ユーザは訪問したスポットに対して殆どの場合高得点をつけていることがわかる．ユーザのレビュー順序が，実際のスポット訪問順序に一致していると仮定すれば，全てのレビューデータを用いて，ユーザ行動データ $\mathcal{D}$ を構築することができる．

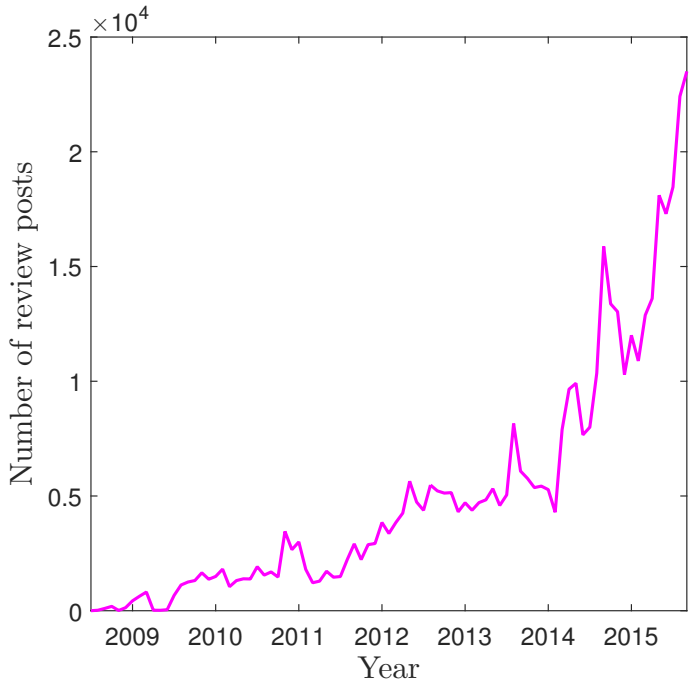

図6.1　レビュー投稿数の推移

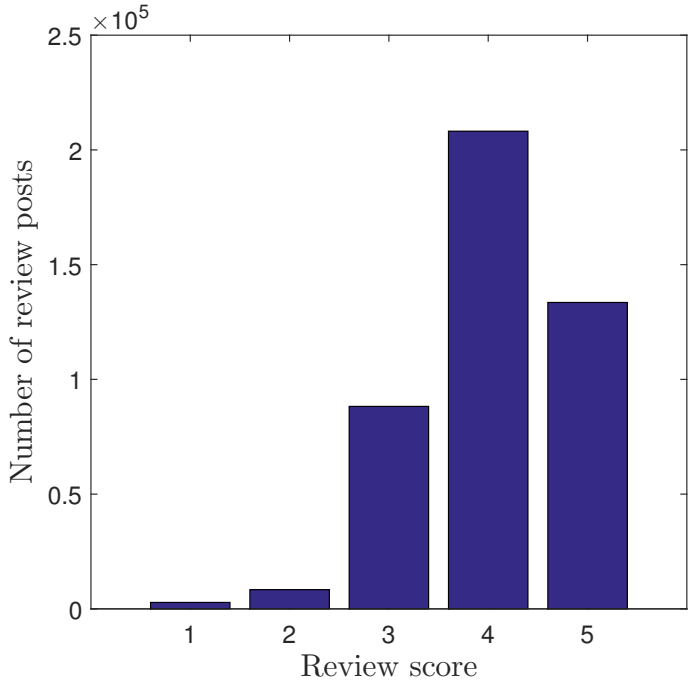

図6.2　レビュー評点の度数分布

続いて，観光スポットの人気度とユーザ行動のスケールフリー性 [45] を検証する．与えられた整数 i に対し，ユーザの度数 $ud(i)$ とスポットの度数 $sd(i)$ を以下のように定義する．

$$ud(i) \ = \ |\{u \in \mathcal{U} \ : \ M(u) = i\}|, sd(i) \ = \ |\{s \in \mathcal{S} \ : \ N(s) = i\}|. \tag{6.9}$$

即ち，$ud(i)$ は i 種類のスポットを訪問したユーザ数を意味しており，$sd(i)$ は i 種類のユーザが訪問したスポット数を意味している．図 6.3，6.4 はユーザとスポットの度数分布を示している．両図から，どちらの度数分布も適度に冪乗則に近似していることがわかるため，スポットの人気度はスケールフリー性を持つという提案手法の仮定は妥当であると言える．図 6.2 から分かるように，ユーザが投稿したレビューの殆どは高評価であり，1 点や2 点のレビューは僅かであるため，スポットに投稿されたレビュー数を，単純にそのスポットの人気度と考えても問題はないように思われる．よってここからは，スポットs に訪れたユーザ数 $N(s)$ を，スポットの人気度 $f(s) = N(s)$ として定義する．

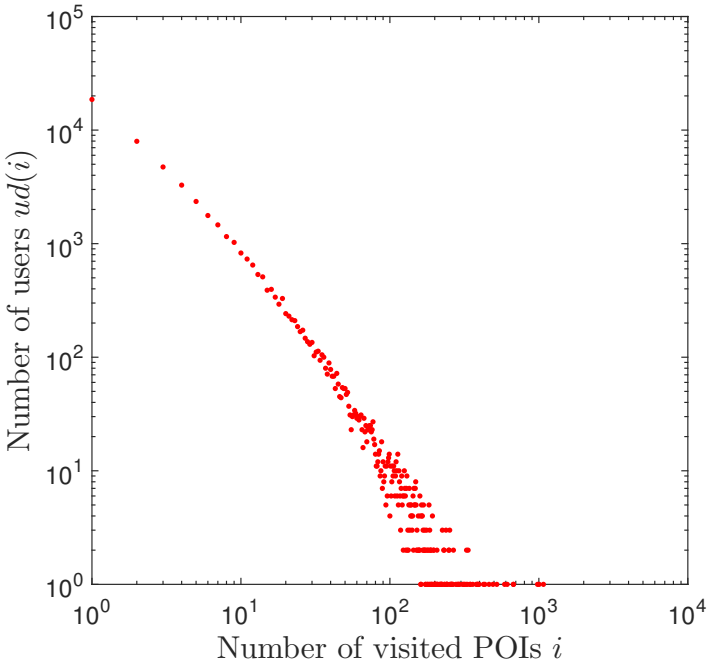

図6.3　ユーザの度数分布

最後に，ユーザ行動データ $\mathcal{D}$ における移動距離のスケールフリー性 [34, 40] を検証する．与えられた距離 Δ に対し，移動距離の度数 $dd(\Delta)$ を以下のように定義する．

$$dd(\Delta) = |\{\cup_{u \in \mathcal{U}} \cup_{1 \leq m < M(u)} (u, m) \ : \ \Delta \leq d(s(u, m), s(u, m+1)) < \Delta + \epsilon\}|. \tag{6.10}$$

スポット間の距離は，スポットの緯度と経度を用いて，GRS80 [31] に基づく測地系によって算出しており，距離間隔 ϵ は 1 km とした．図 6.5 は移動距離の度数分布を示している．移動距離の度数分布

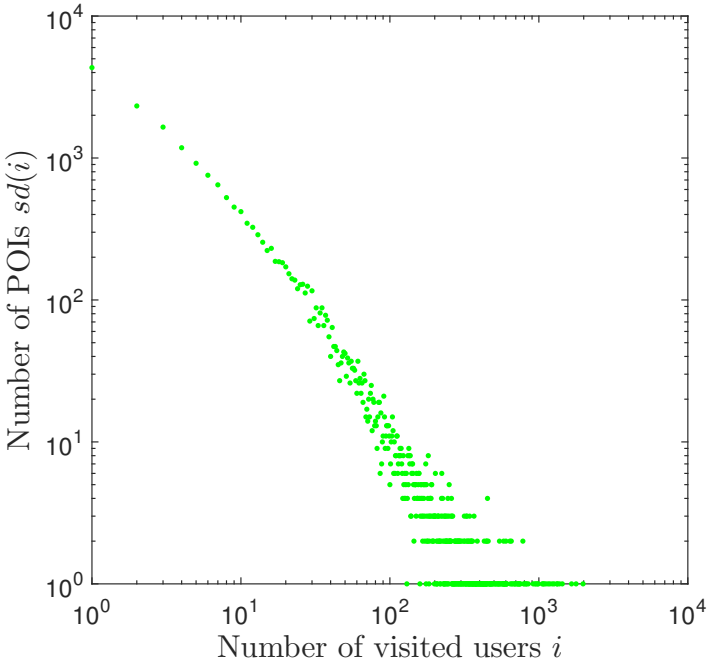

図6.4　スポットの度数分布

も，冪乗則に近似しているため，Lévy flight を仮定している提案手法の妥当性は高いと言える．今回，提案モデルにおける条件付き確率 $p_1(s \mid r; \theta_1)$ が，殆ど距離が離れていないスポット間に対して極めて有効に働くため，レビュー数が多い順に各スポットから半径 100 m を探索していき，探索範囲内に存在する近接スポットのレビューを，探索元スポットのレビューと統合する処理を行った．この処理により，対象スポット数は最終的に $M = 14,319$ となった．

6.5　実験結果

6.5.1　パラメータ推定結果

まず，訪問スポット数 $M(u)$ 毎にユーザを選別し，パラメータ θ_1, θ_2 を推定する．訪問スポット数 $M(u)$ の閾値 τ を導入したときの新たなユーザ行動データ $\mathcal{D}$ は以下となる．

$$\mathcal{D}_\tau = \{(u, v, t) \in \mathcal{D} \ : \ M(u) \geq \tau\}. \tag{6.11}$$

図 6.6 はパラメータの推定結果を示したものであり，横軸は閾値 τ を，縦軸はパラメータ値をそれぞれ表す．図より，θ_1 にはさして変化が見られないが，θ_2 はτ が大きくなるにつれて明らかに減少していることが見て取れる．この結果は，観光スポットを多く訪問するユーザにとっては，スポットの人気度はそれほど重要な要因ではないということを示唆している．この結果の妥当性を裏付けるため，以下の

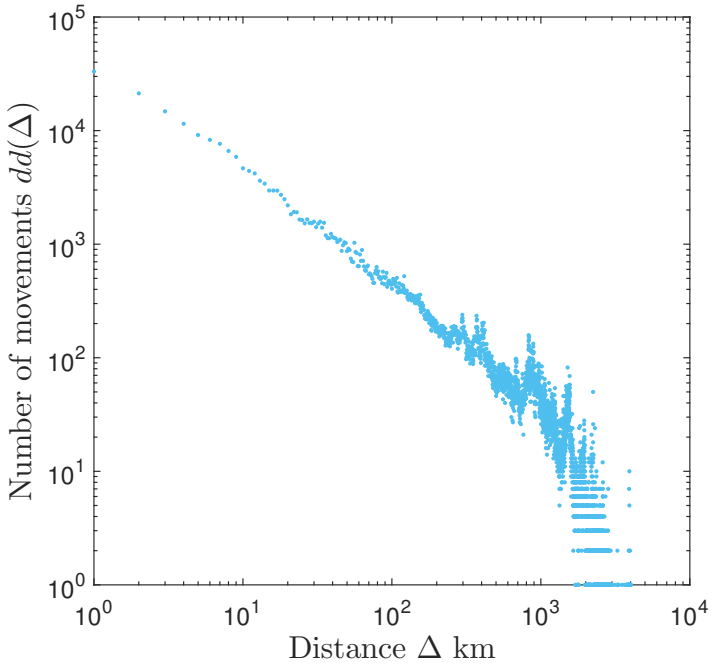

図6.5　移動距離の度数分布

ように定義される次数相関 [46] を考える．

$$dc(i) = \frac{1}{ud(i)} \sum_{\{u \in \mathcal{U} \ : \ M(u)=i\}} \frac{1}{M(u)} \sum_{1 \le m \le M(u)} N(s(u,m)). \tag{6.12}$$

ここで，$N(s)$ はスポット s に訪れたユーザ数であり，スポットの人気度であることに注意されたい．図 6.7 は次数相関の検証結果を示しており，横軸はユーザが訪れたスポット数を，縦軸はユーザが訪れたスポットの人気度の平均をそれぞれ表す．図より，次数相関の観点から見ても，図 6.6 と同様のことが示唆される．即ち，比較的少数の観光スポットしか訪問していないユーザは，概して比較的人気度が高い観光スポットに訪れるということがわかる．

6.5.2　ランキング結果

次に，PageRank 法によるランキングの特性を，人気度によるランキングと入次数法によるランキングとの比較から考察する．いま，PageRank 法によるランキングの上位 k スポットを$R(k)$ とし，人気度と入次数法も同様に，それぞれ$R_{pop}(k)$, $R_{ind}(k)$ とする．そして，以下のランキング類似度 $rs(k;x)$ を用いて比較評価を行う．

$$rs(k;x) = \frac{|R(k) \cap R_x(k)|}{k}. \tag{6.13}$$

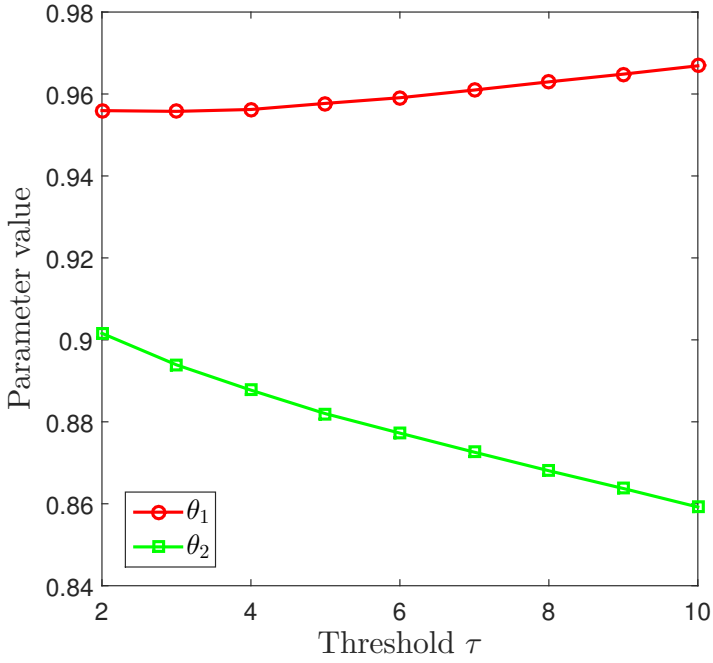

図6.6　閾値 τ 毎のパラメータ推定結果

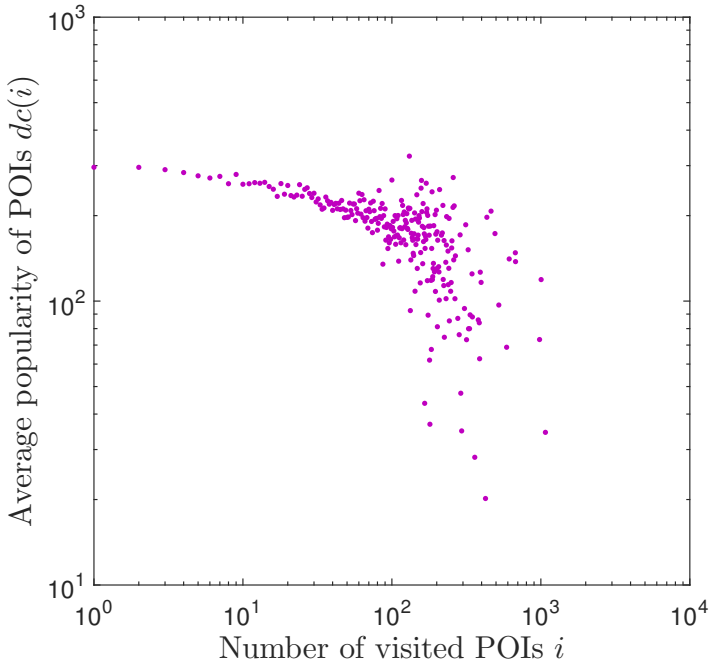

図6.7　次数相関の散布図

ここで，$x \in \{pop, ind\}$ である．図 6.8 はランキング類似度$rs(k;x)$ による評価結果を$k = 1,000$ まで示したものである．図から，人気度とのランキング類似度は平均して 0.70 程度，入次数法とのランキング類似度は平均して 0.80 程度であることがわかる．即ち，これらの手法はそれぞれ，実質的に異なるランキングを生成していると言える．図 6.9, 6.10, 6.11 は，上位 1000 位のスポットを順位で色付けしてプロットし，ランキング結果を可視化したものである．図 6.9 は人気度，図 6.10 は入次数法，図 6.11 はPageRank 法による可視化結果をそれぞれ表す．人気度の上位は比較的日本全体に散らばっているが，PageRank 法の上位は一部の地域に限定して分布しており，入次数法の上位はそれらの中間に位置することが見て取れる．これらの実験結果から，PageRank 法によるランキングは，他の2 手法によるランキングの上位を多く含むいくつかの地域内のスポットに対して集中的に上位を与えていると考える事ができる．

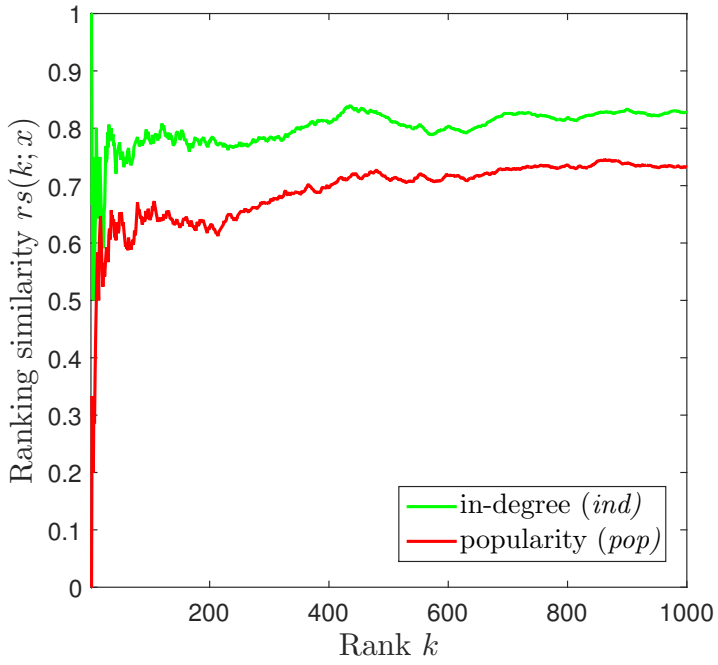

図6.8　ランキング類似度の評価

更に詳細な考察をするため，ランキング結果を用いた地域毎の評価を行う．与えられたスポット集合 S の地域集合を J とする．ここで，地域集合の要素数は $J = |\mathcal{J}|$ とし（スポット集合は $M = |\mathcal{S}|$ ），各要素は整数と同一視されるとする．つまり，$\mathcal{S} = \{1, \cdots, s, \cdots, S\}$ および $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$ である．また，スポット s が属する地域を $j = f(s)$ で表し，各地域に属するスポット数を $M_j = |\mathcal{S}_j| = |\{s\ ;\ j = f(s)\}|$ とする．各スポット s に対し，そのランキングは $1 \leq r_s \leq M$ で与えられているとする．ただし，同順位が起こるケースでは，r_s は平均順位で補正されるとする．ここでの目的

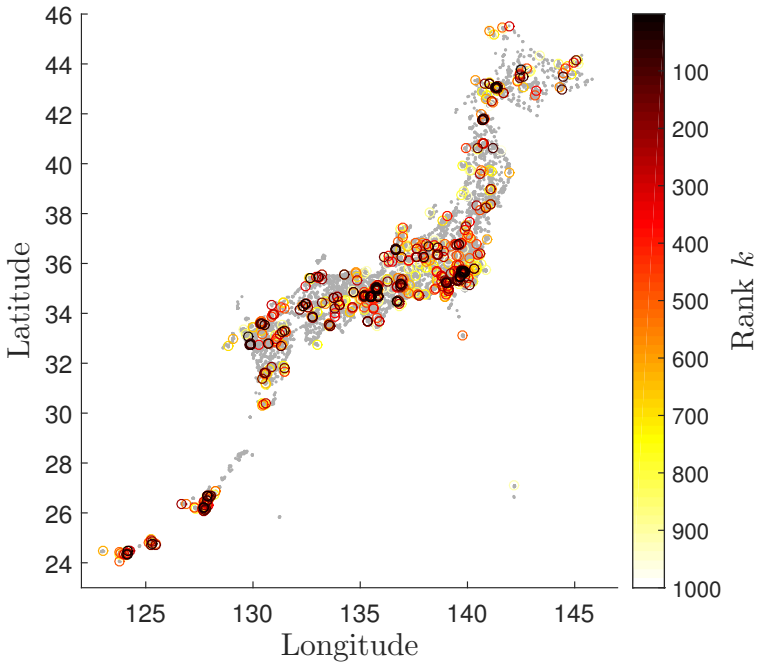

図6.9　人気度ランキングの可視化結果

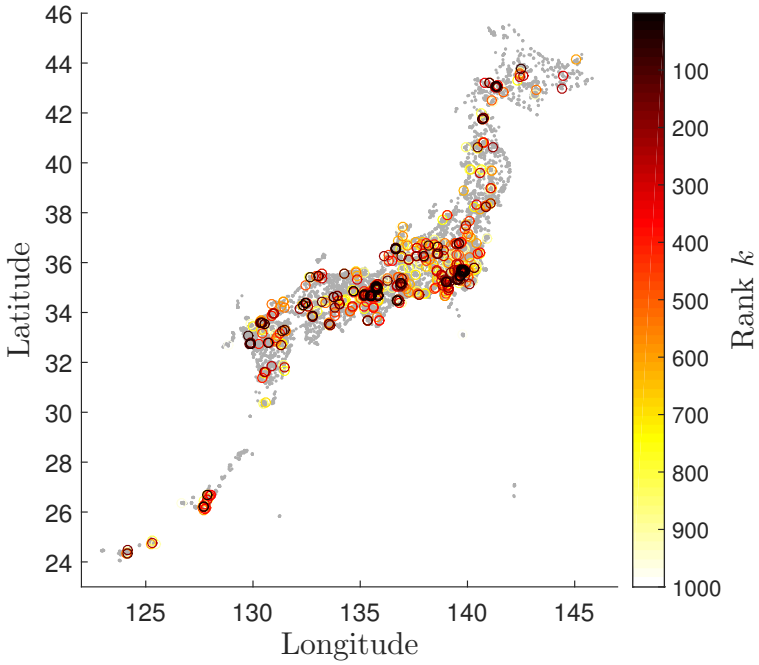

図6.10　入次数法ランキングの可視化結果

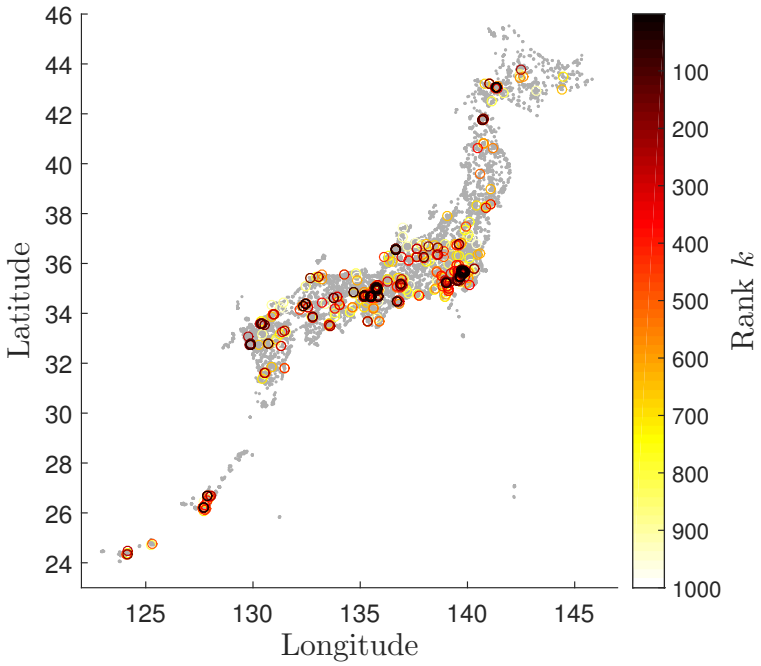

図6.11　PageRank 法ランキングの可視化結果

は，地域集合とランキング付きのスポット集合が与えられたとき，ランキングの高い，または逆に低いスポットが有意に多く含まれる地域を定量的に評価する指標の構築である．以下には 5.4節と同様の，Mann-Whitney の統計量 [32] に基づく自然な拡張法を示す．

Mann-Whitney の2 群順位統計量を多群に拡張して適用する方法について述べる．いま，地域 j に着目すれば，この地域に属すスポット集合 $\mathcal{S}_j$ と，それ以外のスポット集合 $\mathcal{S} \setminus \mathcal{S}_j$ の2 群に分割することができる．ここで，$\cdot \setminus \cdot$ は集合差を意味する．よって，Mann-Whitney の2 群順位統計量に従い，次式により，地域 j に対しz-score z_j を求めることができる．

$$z_j = \frac{u_j - \mu_j}{\sigma_j}, \tag{6.14}$$

$$u_j = M_j(M - M_j) + \frac{M_j(M_j + 1)}{2} - \sum_{s \in \mathcal{S}_j} r_s, \tag{6.15}$$

$$\mu_j = \frac{M_j(M - M_j)}{2}, \tag{6.16}$$

$$\sigma_j^2 = \frac{M_j(M - M_j)(M + 1)}{12}. \tag{6.17}$$

ただし，同順位が起こるケースでは，標準偏差 σ_j は標準的な方法で補正されるとする．よって，式 6.14 で求まるz-score z_j により，各地域 j がランキングの高い，または逆に低いスポットを有意に多く含むか定量的に評価することができる．

まず，地域集合 J を八地方区分としたときの z-score z_j を図 6.12 に示す．北海道地方と九州地方は，人気度のスコアが他の地方よりも明らかに高いが，入次数のスコアは 0 に近く，PageRank のスコアに至っては負値となっている．よって，この 2 地方は，観光地としての人気はあるものの，あくまで一時的に滞在する地域であるということが推察される．それとは逆に，PageRank のスコアが正値となっている関東地方と近畿地方は，観光地というより，拠点地としての役割が強いことが分かる．一方，東北地方，中国地方，四国地方は，人気度のスコアが軒並み低く，入次数のスコアと PageRank のスコアは更に低くなっている．この 3 地方は，観光地として人気が低いだけでなく，拠点地としての役割も弱いことがうかがえる．また，中部地方は，どのスコアにおいても平均的であり，強いて言えば入次数のスコアが比較的高いため，拠点と拠点を繋ぐ中継地点としての役割が強いように思える．

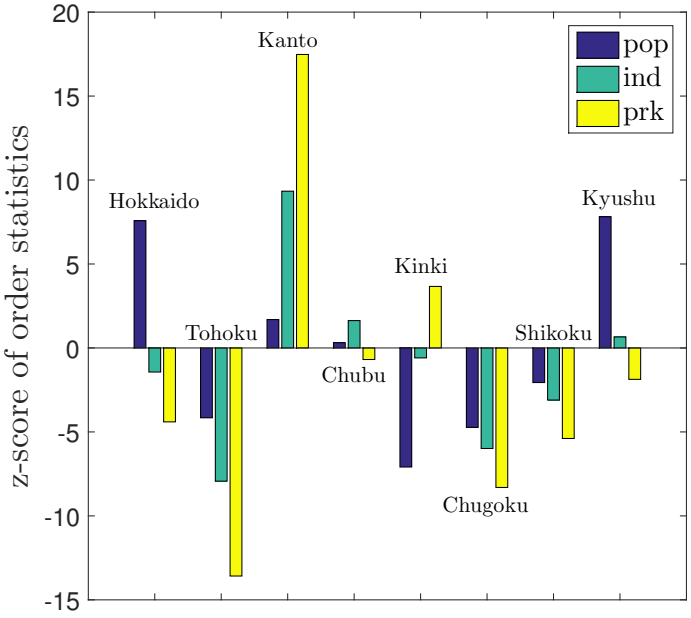

図6.12　八地方区分としたときの各ランキングの順位統計量 z-score

続いて，地域集合 J を都道府県区分としたときの，各ランキングにおける順位統計量の z-score 上位 10 都道府県を表 6.1, 6.2, 6.3 に示す．表 6.1 より，人気度ランキングにおいては，沖縄県が頭一つ抜けてスコアが高く，次いで北海道，東京都が高くなっているため，ここでの上位の地域は，観光地としての人気の高さがうかがえる．表 6.2 より，入次数法ランキングにおいては，東京都が 1 位に上がっており，京都府や愛知県も上位に入っているため，人の流れが多いと思われる地域の z-score が増加していることが見て取れる．表 6.3 より，PageRank 法ランキングにおいては，東京都が突出してスコアが高く，次いで神奈川県，京都府が高くなっており，大阪府も新たに出現していることから，ここでの上

位の地域は，日本全国の観光スポットネットワークにおける優秀な拠点地であると言える．

表6.1　人気度ランキングにおける順位統計量のz-score 上位 10 都道府県

Rank	z_j	Prefucture
1	14.4770	Okinawa
2	7.5794	Hokkaido
3	5.3568	Tokyo
4	2.6809	Kanagawa
5	2.6258	Shizuoka
6	2.2201	Kagoshima
7	2.1930	Nagano
8	1.9683	Chiba
9	1.6848	Ishikawa
10	1.5340	Oita

表6.2　入次数法ランキングにおける順位統計量のz-score 上位 10 都道府県

Rank	z_j	Prefucture
1	12.6052	Tokyo
2	6.5979	Okinawa
3	6.2236	Kanagawa
4	4.4648	Kyoto
5	3.0965	Nagano
6	2.9030	Shizuoka
7	2.2823	Aichi
8	2.1821	Yamanashi
9	1.7685	Chiba
10	1.0773	Ishikawa

6.6　おわりに

本章では，第 5章で述べた時空間モデルの発展系として，観測されたデータのネットワークモデル化を試みた．モデル化実現のために，各観光スポットの人気度を導入した Lévy flight に基づいた確率モデルと，観測されたユーザ行動データからモデルのパラメータを推定する効率的な学習アルゴリズムを提案した．また，提案したモデルから得られた条件付き確率を用いて，2 種類のスポットランキング手法を提案した．レビューサイトのデータセットから生成したユーザ行動データを用いた実験では，パラ

表6.3　PageRank 法ランキングにおける順位統計量のz-score 上位 10 都道府県

Rank	z_j	Prefucture
1	22.0753	Tokyo
2	10.1857	Kanagawa
3	9.4279	Kyoto
4	8.7675	Okinawa
5	3.1010	Chiba
6	2.8854	Shizuoka
7	2.6034	Osaka
8	2.0494	Nagano
9	1.9495	Yamanashi
10	1.8663	Ishikawa

メータ推定に関する詳細な検証と，提案したランキング手法とナイーブな人気度ランキングとの比較を行った．実験結果として，パラメータ推定結果は直感的に解釈が可能であることが示され，提案ランキング手法は地域カテゴリ毎の特性を見出す指標として有用であることが示された．

第7章

カテゴリ評価法：ネットワークモデルの応用

本章では，第 6章のネットワークモデルの応用について述べる．例えば，位置情報が無かったり，時刻情報もバッチ処理的に統一されているようなデータにおいては，ネットワーク内の移動という概念は使えない．しかし，データオブジェクトをある種のカテゴリに分けたときのカテゴリ間類似度は，大抵のデータにおいて計算することができるため，それを使った巨視的かつ汎用的なネットワークモデルを提案する．応用ネットワークモデルでは，カテゴリがネットワークのノードとなるため，第 6章で提案したノードの評価指標そのものがカテゴリ評価指標として用いることができる．ここでは，それらカテゴリ評価指標の時系列的変化に着目し，クラスタリングによってカテゴリに関する規則性や重要性を見出すことを試みる．本章では，近年のインターネット上の代表的なカテゴリとしてソーシャルタグを扱う．

なお，数式や解法においては，本章で全て定義し直すものとする．

7.1　はじめに

ソーシャルメディアの発展により，近年，様々な Web オブジェクトに対してソーシャルタグが付与されるようになった．インターネット上のソーシャルタグは，ユーザの個々の意思決定に基いて常に生成され続けているため，各ソーシャルタグの役割や機能が，時間とともに変化することは自然と推察される．よって本章では，巨視的分析の見地として，ソーシャルタグ間の確率ネットワークを構築し，それを分析することでソーシャルタグに関する規則性や重要性を見出すことを試みる．より詳細には，まず，ソーシャルタグ間の類似度を確率として正規化し，それらを確率有向リンクとしてソーシャルタグ間の確率ネットワークを構築する方法を提案する．次に，それら条件付き確率に基づいた各ノードの PageRank 値を高速に求める計算法を提案し，それを用いたソーシャルタグの分類手法を提案する．提

案分析手法の評価には，ニコニコ動画 [*1]のデータセットから生成した，タグ共起データを用いる．

7.2　提案分析手法

7.2.1　確率ネットワーク生成法と PageRank 値計算法

与えられたソーシャルタグ集合と，ニコニコ動画などタグが付与されるオブジェクト集合のそれぞれを自然数と同一視し，$\mathcal{M} = \{1, \cdots, m, \cdots, M\}$ と $\mathcal{H} = \{1, \cdots, h, \cdots, H\}$ で表す．ここで，$M = |\mathcal{M}|$ と $H = |\mathcal{H}|$ は総タグ数と総オブジェクト数である．また，時刻 t の時点でタグ m が付与されていたオブジェクト集合を $\mathcal{H}_m^{(t)} \subset \mathcal{H}$ とする．ただし，時刻 t も自然数と同一視して $t \in \{1, \cdots, T\}$ とし，T は最終観測時刻を表すとする．さらに，オブジェクト集合 $\mathcal{H}_m^{(t)}$ より，$h \in \mathcal{H}_m^{(t)}$ ならば $x_{m,h}^{(t)} = 1$ とし，さもなければ $x_{m,h}^{(t)} = 0$ として，各タグ m に対して H-次元縦ベクトル $\mathbf{x}_m^{(t)}$ を定義する．ただし，時刻の指定が不要な場合には，$\mathbf{x}_m^{(t)}$ を $\mathbf{x}_m$ と略記する．以降では，ベクトルとして定義するものは全て縦ベクトルとする．

いま，任意の時刻 t におけるソーシャルタグ間の $M \times M$ 類似度行列 $\mathbf{S}^{(t)}$ （時刻の指定が不要な場合には $\mathbf{S}$ ）をコサイン類似度に基づき定義する．すなわち，任意のタグのペア $n, m \in \mathcal{M}$ に対し，$\mathbf{x}_m \neq \mathbf{0}_H$ かつ $\mathbf{x}_n \neq \mathbf{0}_H$ ならば，$S(m,n) = (\mathbf{x}_m^T \mathbf{x}_n)/(||\mathbf{x}_m||||\mathbf{x}_n||)$ とし，さもなければ $S(m,n) = 0$ とする．ただし，$\mathbf{0}_H$ は H-次元の 0 ベクトルであり，$\mathbf{x}_m^T$ はベクトル $\mathbf{x}_m$ の転置を表し，$||\mathbf{x}_m||$ は $\sqrt{\mathbf{x}_m^T \mathbf{x}_m}$ で定義されるベクトル $\mathbf{x}_m$ のノルムである．次に，この類似度行列 $\mathbf{S}$ を正規化して，任意の時刻 t における $M \times M$ 推移確率行列 $\mathbf{P}^{(t)}$ （時刻の指定が不要な場合には $\mathbf{P}$ ）を構成する．ここで，類似度行列 $\mathbf{S}$ の要素 $S(m,n)$ を用いて，第 m 行の和を $S(m) = S(m,1) + \cdots + S(m,M)$ で定義する．推移確率行列 $\mathbf{P}$ の要素 $P(m,n)$ は，$S(m) \neq 0$ かつ $m \neq n$ ならば，$P(m,n) = S(m,n)/S(m)$ とし，$S(m) \neq 0$ かつ $m = n$ ならば，$P(m,n) = 0$ とし，そして $S(m) = 0$ ならば $P(m,n) = 1/M$ とする．すなわち，推移確率行列 $\mathbf{P}$ は，自己リンクなしで，タグ m から類似が高いタグに高い確率で推移し，$\mathbf{x}_m = \mathbf{0}_H$ のときなど任意のタグとの類似が 0 の場合は任意のタグへのランダムな推移となる．

推移確率行列 $\mathbf{P}$ に対し，一様ジャンプ確率 $\alpha \in (0,1)$ を用いて，任意の時刻 t における Google 行列 $\mathbf{G}^{(t)}$ （時刻の指定が不要な場合には $\mathbf{G}$）を $\mathbf{G} = (1-\alpha)\mathbf{P} + \alpha \mathbf{1}_M \mathbf{1}_M^T$ で定義すれば，ソーシャルタグ間の確率ネットワークにおいて，各タグの PageRank 値を求めることができる．ここで，$\mathbf{1}_M$ は任意の要素値が 1 の M-次元ベクトルを表す．しかしながら，任意の時刻 t で Google 行列 $\mathbf{G}^{(t)}$ を求めて，各タグの PageRank 値の時系列を求めるとすれば，ある時刻 t の類似度行列 $\mathbf{S}$ を求めるのに $O(M^2 H)$ の計算量が必要となり，それを時刻 $t = 1$ から最終観測時刻 T まで求めるため，全体で $O(TM^2 H)$ の計算量が必要となる．よって，この計算量では大規模データへの適用は困難な場合も起こる．

[*1] www.nicovideo.jp

以下では，確率ネットワークでの PageRank 値を高速に求める計算法を提案する．いま，$S(m) \neq 0$ となるタグ数が $N(\leq M)$ のとき，$m \leq N$ ならば $S(m) \neq 0$ となり，$m > N$ ならば $S(m) = 0$ となるように，$\mathcal{M}$ の要素を並び替えても一般性を失わない．また，$S(m) \neq 0$ となるタグに対し，これらベクトル $\mathbf{x}_m$ を並べて構成する $H \times N$ 行列を $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$ とし，これらの正規化値 $S(m)$ を要素とする $N \times N$ 対角行列を $\mathbf{\Delta} = diag(S(1), \cdots, S(N))$ とし，そして，これらベクトルのノルム $||\mathbf{x}_m||$ を要素とする$N \times N$ 対角行列を $\mathbf{\Gamma} = diag(||\mathbf{x}_1||, \cdots, ||\mathbf{x}_N||)$ とする．このとき，推移確率行列 $\mathbf{P}'$ は以下のように表せる．

$$\mathbf{P}' = \begin{pmatrix} \mathbf{\Delta}^{-1}\left(\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{\Gamma}^{-1} - \mathbf{I}_{N,N}\right), & \mathbf{0}_{N,M-N} \\ \frac{1}{M}\mathbf{1}_{M-N}\mathbf{1}_M^T & \end{pmatrix}.$$

ここで，$\mathbf{I}_{N,N}$ は $N \times N$ 単位行列であり，$\mathbf{0}_{N,M-N}$ は全ての要素が 0 の $N \times (M-N)$ 行列を表す．いま，各タグの PageRank 値ベクトルを $\mathbf{y}^T = (\mathbf{u}^T, \mathbf{v}^T)$ とする．ここで，$\mathbf{u}$ は $S(m) \neq 0$ となるタグに対する N-次元ベクトルであり，$\mathbf{v}$ は $S(m) = 0$ となるタグに対する $(M-N)$-次元ベクトルである．よって， Google 行列 $\mathbf{G}$ の定義に従えば，ベクトル $\mathbf{y}$ から PageRank 更新式より求まる次のステップのベクトル $\bar{\mathbf{y}}$ は以下となる．

$$\begin{aligned} \bar{\mathbf{y}}^T = & \left((1-\alpha)\mathbf{u}^T\mathbf{\Delta}^{-1}\left(\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{\Gamma}^{-1} - \mathbf{I}_{N,N}\right), \mathbf{0}_{M-N}\right) \\ & + \frac{\mathbf{v}^T\mathbf{1}_{M-N} + \alpha}{M}\mathbf{1}_M^T. \end{aligned} \tag{7.1}$$

明かに，行列 $\mathbf{\Delta}^{-1}$ や $\mathbf{\Gamma}^{-1}$ との積は，それぞれ対角行列なので，$O(N)$ 回の乗算で求まる．一方，行列 $\mathbf{X}$ の任意の要素は 0 または 1 であり，1 の要素数は各オブジェクトに付与されたタグ数の合計で，$L = \mathbf{1}_H^T\mathbf{X}\mathbf{1}_M$ となることより，行列 $\mathbf{X}^T$ や $\mathbf{X}$ との積は高々 L 回の加算で求まる．したがって，式 (7.1) の更新は，$O(N)$ 回の乗算と $2L$ 回の加算で実現できる．以下に提案法のアルゴリズムを示す．

1. PageRank 値ベクトルを $\mathbf{y} = (1/\sqrt{M}, \cdots 1/\sqrt{M})^T$ と初期化する；
2. 式 (7.1)で PageRank 値ベクトルを $\bar{\mathbf{y}}$ を求める；
3. $\sum_{m\in\mathcal{M}} |y_m - \bar{y}_m| < \epsilon$ ならば $\bar{\mathbf{y}}$ を出力し終了する；
4. $\mathbf{y} \leftarrow \bar{\mathbf{y}}$ としステップ2.へ戻る．

実験では， Google 行列を構成するための一様ジャンプ確率を $\alpha = 0.15$ とし，終了条件を制御するパラメータを $\epsilon = 10^{-8}$ に設定した．なお，推移確率行列 $\mathbf{P}$ を陽に求め，上記ステップ2. を $\bar{\mathbf{y}}^T \leftarrow \mathbf{y}^T\mathbf{P}$ で求める方法をベースライン法と呼び，実験では，ベースライン法との比較により，提案法の性能を評価する．

7.2.2　k-medoids クラスタリング

k-medoids 法は，非階層クラスタリングで有名な k-means 法と同様に，H-次元の M 個のオブジェクト集合 $\mathcal{M}$ が与えられたとき，オブジェクト集合を K 個のクラスタに分割する問題を解くための手法である．任意のオブジェクトペア $u, v \in \mathcal{M}$ 間に類似度 $\rho(u, v)$ が定義されていれば，オブジェクト集合の中から他のオブジェクトとの類似度の和が高い代表オブジェクトを選定することが可能であるため，最適な代表オブジェクトが選定されれば，類似度の高いオブジェクトペアは同じクラスタに，類似度の低いオブジェクトペアは異なるクラスタに属するように分割されるはずである．このような問題では，一般的に平均（mean）より中央値（median）の方が頑健であることが知られている．ただし，大域最適解を求めるためには $O(M^K H)$ の計算量が必要であるため，オブジェクト集合の規模や次元，分割数 K がある程度大きくなると，実用的な時間で解を求めることが難しくなる．よって，k-medoids 法にも局所最適解を求めるための反復法や貪欲法が存在するが，今回は解の一意性が保証される貪欲法に基づく解法を採用する．この解法は，目的関数のサブモジュラ性により，厳密解ではないものの，ある程度妥当な精度で最悪ケースの解品質が理論的に保証されている [47].

貪欲法とは，既に選定した代表オブジェクトを固定し，目的関数値を最大にするオブジェクトを求め，目的関数が増加するならば代表オブジェクト集合に追加することで，結果の代表オブジェクト集合を求める方法である．各オブジェクトは，最も類似度の高い代表オブジェクトと同じクラスタに割り当てられる．既に選定した代表オブジェクト集合を $\mathcal{P}$ とし，新たに追加を試みるオブジェクトを w とするとき，ここでは，以下の目的関数を考える．

$$f(\mathcal{P} \cup \{w\}) = \sum_{v \in \mathcal{M}} \max\{\mu(v\,; \mathcal{P}), \rho(v, w)\}. \tag{7.2}$$

ここで，$\mu(v\,; \mathcal{P})$ は既に選定された代表オブジェクトとの類似度の最大値を表し，$\mu(v\,; \mathcal{P}) = \max_{w \in \mathcal{P}}\{\rho(v, w)\}$ で定義される．以下に k-medoids 法における貪欲法アルゴリズムを説明する．ここで，$\cdot\backslash\cdot$ は集合差を表す．

A1-1. $k \leftarrow 1$, $\mathcal{P}_0 \leftarrow \emptyset$，各オブジェクト $v \in \mathcal{M}$ に対し，$\mu(v\,; \emptyset) \leftarrow 0$ と初期化する；

A1-2. 式 (7.2)で $\hat{p}_k = \mathrm{argmax}_{w \in \mathcal{M} \backslash \mathcal{P}_{k-1}}\{f(\mathcal{P}_{k-1} \cup \{w\})\}$ を求め，$\mathcal{P}_k \leftarrow \mathcal{P}_{k-1} \cup \{\hat{p}_k\}$ とする；

A1-3. $k = K$ ならば $\hat{\mathcal{P}}_K = \{\hat{p}_1, \ldots, \hat{p}_K\}$ を出力し終了する；

A1-4. 各オブジェクト $v \in \mathcal{M}$ に対し，$\mu(v\,; \mathcal{P}_k)$ を求める；

A1-5. $k \leftarrow k + 1$ とし，ステップ A1-2 へ戻る．

各オブジェクトを，最も類似度の高い代表オブジェクト $p_k \in \mathcal{P}$ のクラスタ $\mathcal{C}_k$ に割り当てる．明らかに，上記のアルゴリズムの計算量は $O(M^2 K H)$ となるため，大域最適解を得るために必要な計算量

$O(M^K H)$ に比べて非常に高速である．しかし，貪欲法に基づく単純な手法であるため，比較的プアーな局所解にトラップされる危険性が伴う．

よって，以下では貪欲法アルゴリズムで得た $\hat{\mathcal{P}}_K$ の解品質を向上させるための局所探索法アルゴリズムについて述べる．

A2-1. $k \leftarrow 1, h \leftarrow 0$ と初期化する；

A2-2. 式 (7.2)で

$p'_k = \mathrm{argmax}_{w \in \mathcal{M} \backslash \hat{\mathcal{P}}_K \backslash \{\hat{p}_k\}} \{f(\hat{\mathcal{P}}_K \backslash \{\hat{p}_k\} \cup \{w\})\}$ を求める；

A2-3. $p'_k = \hat{p}_k$ ならば $h \leftarrow h+1$ とし，さもなければ $h \leftarrow 0, \hat{\mathcal{P}}_K = \hat{\mathcal{P}}_K \backslash \{\hat{p}_k\} \cup \{p'_k\}$ とする；

A2-4. $h = K$ ならば $\hat{\mathcal{P}}_K$ を出力して終了する；

A2-5. 各オブジェクト $v \in \mathcal{M}$ に対し，$\mu(v\ ; \hat{\mathcal{P}}_K)$ を求め，$k = K$ ならば $k \leftarrow 1$，さもなければ $k \leftarrow k+1$ とし，ステップ A2-2 へ戻る．

第 2章から第 4章まで扱った区間分割法に関する種々の実験結果より，最適化問題における貪欲法と局所探索法の反復的手法は，計算量の増加こそあるものの，解精度を向上させる効果が期待できるということが既に分かっている．先にも述べたように，このクラスタリング手法には解の一意性が保証されているため，基本的にはこの計算処理は一度しか行われない．よって，多少の計算量の増加を負ってでも，一度で解精度が高い結果を出すために，ここでは貪欲法アルゴリズムと局所探索法アルゴリズムを反復して使用する手法を述べる．

I1. A1-1 から処理を開始する；

I2. A1-4 の処理後に $k > 1$ ならば $\mathcal{P}_k$ を $\hat{\mathcal{P}}_K$ として出力する；

I3. $\hat{\mathcal{P}}_K$ を A2 で改善し，改善した $\hat{\mathcal{P}}_K$ を $\mathcal{P}_k$ として出力する；

I4. A1-5 から処理を再開させ，ステップ I2 へ戻る．

7.3　データセット

今回の実験で用いるデータセットは，ニコニコ動画におけるVOCALOIDオリジナル楽曲動画のタグ [*2]の時系列データである．このデータセットは，VOCALOIDオリジナル楽曲動画が一般に有するタグによる検索結果から，二次創作系や加工系のタグを有する動画を除外して取得したものであり，取得

[*2] 1つの動画につき11個まで登録できる関連文字列

期間は 2013/04/03 から 2016/03/16 (24時間毎 1078 日間 $T = 1078$)，最終日の動画数 H は 136192，タグ数 M は 125196 である．今回，取得期間中に一度でも 10 以上の動画に登録されていたタグを分析対象としたため，最終的に使用したタグの種類 M は 7787 である．

ここでは，データセットに対して行った統計的処理の結果について述べる．図 7.1，7.2は，それぞれ取得期間中の動画数と分析対象タグ数の推移であり，どちらも一定のペースで増加し続けていることが分かる．これにより，時間の経過によるタグネットワークの構造変化が期待できる．図 7.3は，最終観測時刻における全てのタグを用いたときのタグの度数分布である．多くの Web オブジェクトと同様のように，ニコニコ動画のタグの使用頻度にもスケールフリー性が見られる．

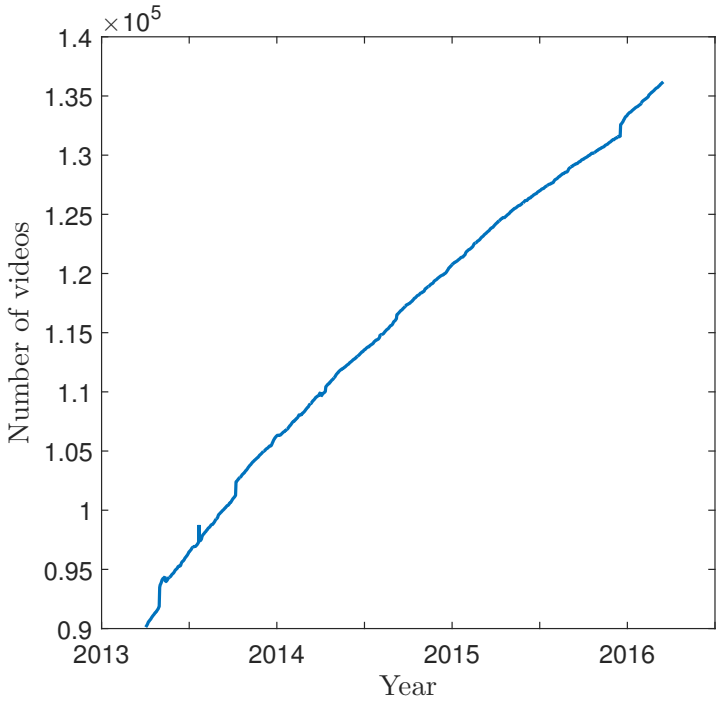

図7.1　動画数の推移

7.4　評価実験

今回の実験結果について述べる．まず，提案手法とベースライン法による PageRank 値の計算時間の比較を図 7.4に示す．図より，ベースライン法は使用されたタグ数の増加と共に計算時間が増加し続けているが，提案手法はどのようなタグ数に対しても計算時間が1秒未満となっているため，提案手法が圧倒的に高速であることは明らかである．

次に，タグ確率ネットワークの期待される性質を調べるため，分析対象タグの共起関係を次数相関に

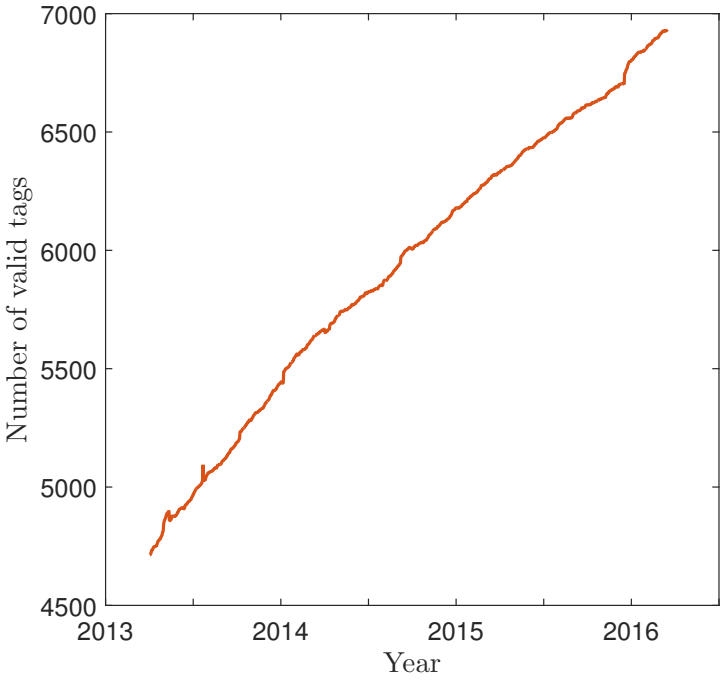

図7.2　対象タグ数の推移

見立ててプロットしたものを図 7.5に示す．図の横軸はタグの出現数を，縦軸は共起しているタグの出現数の平均を表している．図より，出現数が 10 から 100 にかけてのタグは，共起しているタグの出現数平均が減少する傾向があるが，それよりも出現数が多いタグに関しては，全体的に共起しているタグの出現数平均が増加する傾向があるため，生成されるネットワークの特徴としては，次数が大きい（出現数と共起数が共に多い）ノードがリンク確率（他ノードからの確率）を得やすいことが期待される．

ここからは，提案 k-medoids 法によるタグのクラスタリング結果について述べる．今回のデータセットにおける PageRank 値の時系列データを prk ，すなわち

$$prk = \left(\ \bar{\mathbf{y}}^{(1)}, \ \cdots, \ \bar{\mathbf{y}}^{(T)} \ \right),$$

（$\bar{\mathbf{y}}^{(t)}$ は任意の時刻 t において求められた PageRank 値ベクトル）とし，推移確率行列 $\mathbf{P}$ における各ノードの入次数確率合計の時系列データを ind, すなわち

$$ind = \begin{pmatrix} \sum_{m=1}^{M} P^{(1)}(m,1) & \cdots & \sum_{m=1}^{M} P^{(T)}(m,1) \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{M} P^{(1)}(m,M) & \cdots & \sum_{m=1}^{M} P^{(T)}(m,M) \end{pmatrix},$$

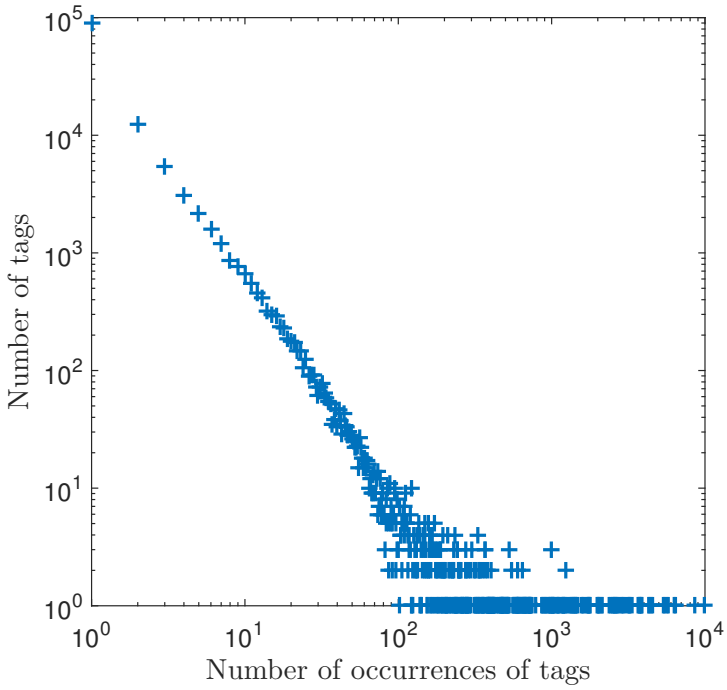

図7.3　タグの度数分布

タグの出現数の時系列データを *pop*，すなわち

$$pop = \begin{pmatrix} \sum_{h=1}^{H} \mathbf{x}_1^{(1)} & \cdots & \sum_{h=1}^{H} \mathbf{x}_1^{(T)} \\ \vdots & \ddots & \vdots \\ \sum_{h=1}^{H} \mathbf{x}_M^{(1)} & \cdots & \sum_{h=1}^{H} \mathbf{x}_M^{(T)} \end{pmatrix},$$

として比較に用いる．なお，時系列データの類似度 $\rho(u, v)$ は相関係数とした．図 7.6 にクラスタ数 K と解品質の関係を示す．図の縦軸は，各クラスタの代表オブジェクトとの類似度の最大値の総和 $\sum_{v \in \mathcal{M}} \mu(v\,;\,\hat{\mathcal{P}}_K)$ を表している．図より，出現数 *pop* に基づくクラスタリングにおけるクラスタの解品質が高く，入次数確率 *ind* に基づくクラスタリングにおけるクラスタの解品質が低いことがわかる．また，K が 20 を超えたあたりでどの指標においても解品質の改善が鈍くなるなるため，以下，分析用のクラスタ数は $K = 20$ とした．図 7.7は，$K = 20$ のときの各クラスタの代表オブジェクトとの類似度の平均を示したものである．図の横軸は類似度の平均が高い順で降順ソートしたときのランクである．クラスタ内の類似度平均が最も高いクラスタは *prk* 内に存在しているが，全体を見るとやはり *pop* のクラスタ品質が高い傾向にある．詳細な検証として，*pop* と *prk* のクラスタ品質上位2クラスタを比較する．まず，クラスタ品質1位の比較を表 7.1,7.2に示す．どちらも似たようなタグが並んでいるため，大きな差異が無いように思えるが，タグの種別頻度で見てみると，*pop* は ‘category’ 3,

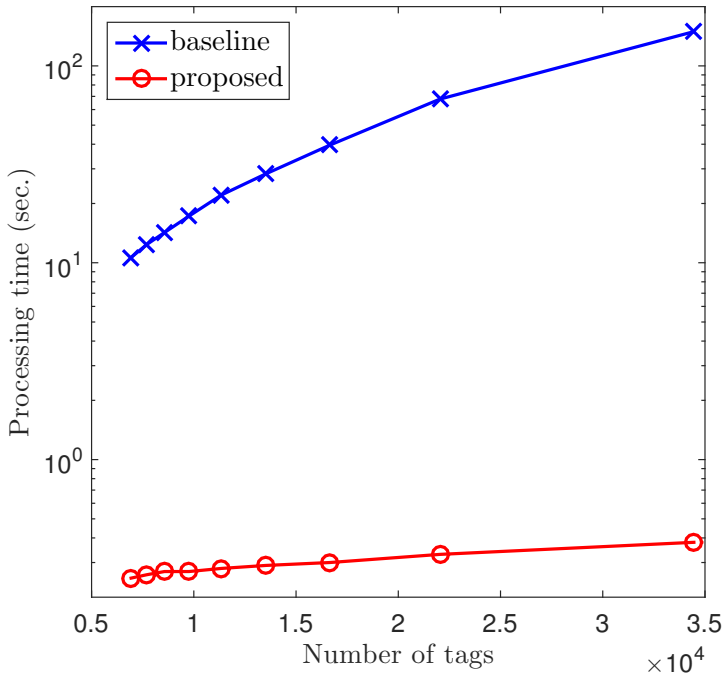

図7.4　PageRank 値の計算時間比較

‘software’ 13, ‘genre’ 3, ‘information’ 1 となっており，*prk* は ‘category’ 2, ‘software’ 15, ‘genre’ 3 となっているため，多少なり *prk* の方がタグの種別のばらつきが小さいことが分かる．続いて，クラスタ品質2位の比較を表 7.3,7.4に示す．ここでも似たようなタグが並んでおり，タグの種別も ‘genre’, ‘artist’, ‘dancer’ が主であるため，目立った差異が無いように見える．恐らく，どちらも Dubstep という音楽ジャンル周辺で見られるタグと思われるが，もし Dubstep が基軸となったクラスタであるなら，*pop* のラインナップはそこまで妥当とは言えない．現に，*pop*の表においては，上位2つを除いて Dubstep に関連しているタグは5位の ‘skrillex’(Dubstep artist) と16位の ‘東方dubstep’(Dubstep genre) の2つだけである．それに対し，*prk*の表においては，4位の ‘skrillex’(Dubstep artist), 10位の ‘東方dubstep’(Dubstep genre), 13位の ‘ukf’(Dubstep artists bland), 14位の ‘klaypex’(Dubstep artist), 15位の ‘nero’(Dubstep artist), 16位の ’ブロステップ‘(Dubstep genre), 18位の ‘skream’(Dubstep artist), 19位の ‘knife party’(Dubstep artist), 20位の ‘脱糞ステップ’(Dubstep genre) の9個のタグが Dubstep と深く関わっている．

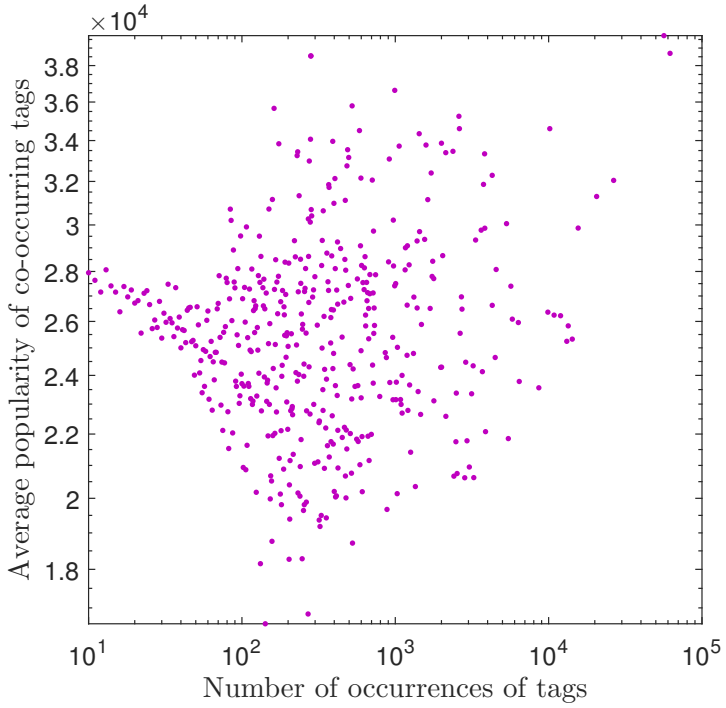

図7.5　タグの共起関係に基づく次数相関

7.5　おわりに

本章では，第 6章のネットワークモデルの応用として，カテゴリ（今回はソーシャルタグ）間の類似度による確率ネットワーク生成法を提案した．さらに，提案確率ネットワークを分析することで，ソーシャルタグに関する規則性や重要性を見出すことを試みた．今回提案した確率ネットワーク生成法と PageRank 値計算法は，ベースライン手法と比較して圧倒的に高速に PageRank 値を算出することができるため，多数の観測時刻におけるデータを PageRank 時系列データとして容易に扱えることを示した．また，提案した k-medoids アルゴリズムによる PageRank 時系列データのクラスタリングでは，ソーシャルタグの役割や機能に即した出力が見られたため，ソーシャルタグの分析手法としての有用性が期待できる．しかし，図 7.8に示すように，提案 k-medoids アルゴリズムはクラスタ数 K の増加による計算時間の増加が著しいため，極端に多いクラスタ数には向いていない．

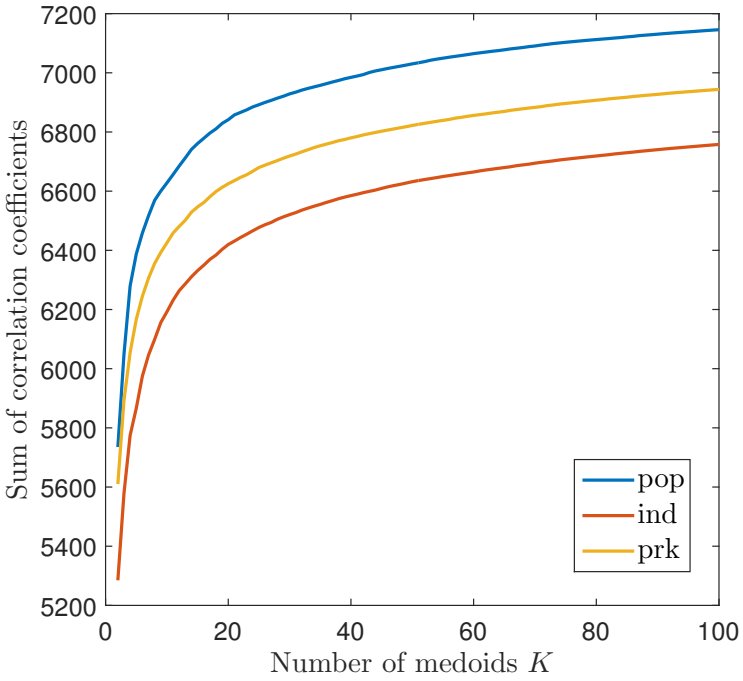

図7.6　クラスタ数 K と解品質の関係

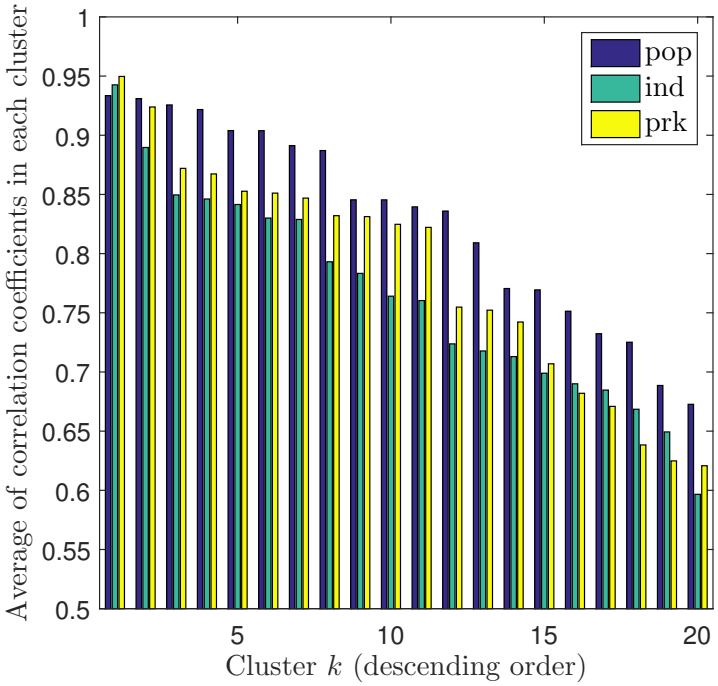

図7.7　クラスタ品質の検証

表7.1　*pop* の品質1位のクラスタ内における出現数上位20タグ

	pop	
rank	type	tag
1	category	vocaloid
2	software	初音ミク
3	software	ミクオリジナル曲
4	genre	vocarock
5	software	鏡音リン
6	software	gumi
7	software	gumiオリジナル曲
8	software	巡音ルカ
9	software	リンオリジナル曲
10	information	vocaloid処女作
11	software	ルカオリジナル曲
12	software	鏡音レン
13	software	レンオリジナル曲
14	software	ia -aria on the planetes-
15	software	kaito
16	software	iaオリジナル曲
17	genre	ボカロクラシカ
18	category	vocaloidオリジナル曲
19	category	vocaloid-pv
20	genre	ボカロバラード

表7.2　*prk* の品質1位のクラスタ内における出現数上位20タグ

	prk	
rank	type	tag
1	category	vocaloid
2	software	初音ミク
3	software	ミクオリジナル曲
4	category	音楽
5	genre	vocarock
6	software	鏡音リン
7	software	gumi
8	software	巡音ルカ
9	software	リンオリジナル曲
10	software	ルカオリジナル曲
11	software	鏡音レン
12	software	レンオリジナル曲
13	software	kaito
14	software	ニコニコムービーメーカー
15	genre	ボカロクラシカ
16	software	メグッポイド
17	software	vocaloid3
18	software	meiko
19	software	mikumikudance
20	genre	爽やかなミクうた

表7.3　*pop* の品質2位のクラスタ内における出現数上位20タグ

	pop	
rank	type	tag
1	genre	dubstep
2	genre	ダブステップ
3	genre	electro
4	genre	洋楽
5	artist	skrillex
6	genre	アニメ色のない作業用bgm
7	dancer	remotekontrol
8	genre	street dance統一タグ
9	genre	amv
10	artist	uk
11	genre	dj
12	genre	nonstop
13	dancer	marquese scott
14	dancer	左のおっさん
15	genre	フリームーブ
16	genre	東方dubstep
17	genre	dnb
18	genre	アングラアニソンremixリンク
19	genre	アニメーションダンス
20	dancer	chibi

表7.4　*prk* の品質2位のクラスタ内における出現数上位20タグ

	prk	
rank	type	tag
1	genre	dubstep
2	genre	ダブステップ
3	genre	洋楽
4	artist	skrillex
5	artist	remotekontrol
6	artist	uk
7	genre	nonstop
8	dancer	marquese scott
9	dancer	左のおっさん
10	genre	東方dubstep
11	genre	アングラアニソンremixリンク
12	dancer	chibi
13	brand	ukf
14	artist	klaypex
15	artist	nero
16	genre	ブロステップ
17	dancer	bryan gaynor
18	artist	skream
19	artist	knife party
20	genre	脱糞ステップ

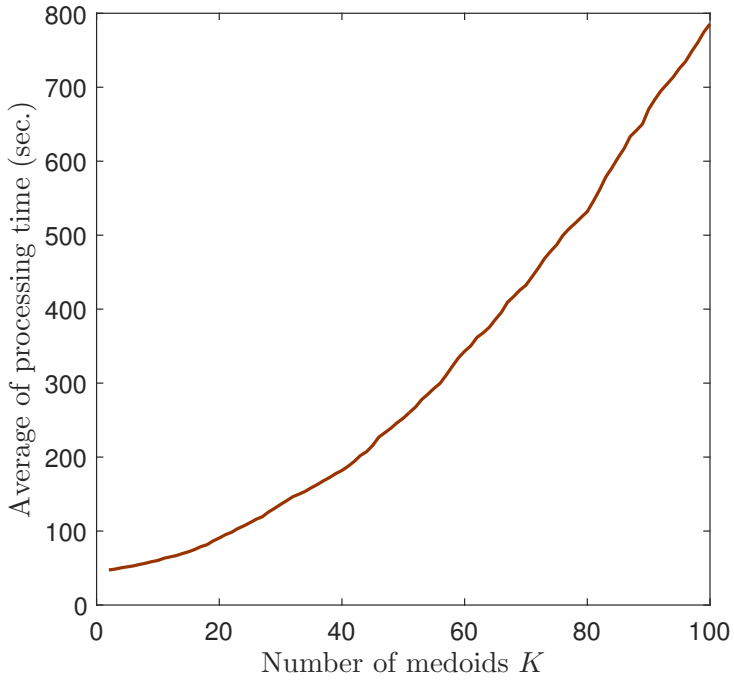

図7.8　貪欲法アルゴリズムと局所探索法アルゴリズムの反復による計算量の増加（3指標による平均）

第8章

結論

本論文では，多様な環境において生成された多種データに対する汎用技術についての一連の研究について述べた．以下に，各章で提案した技術と，それにより得られたものをまとめる．

第 2章では，代表的な異常（バースト）検出手法特有の問題を補うため，属性が複数含まれていたり，観測時刻間隔が一定だったりするような多種データを扱うことを前提として，多様な区間長の異常（バースト）を検出する手法を提案した．ここでは，データ属性の確率分布に対する尤度最大化と尤度比検定に基づいた問題設定を行い，その問題を高速且つ高精度に解くための解法を提案した．人工データを使った Kleinberg の手法との比較実験では，両手法の異常検出に関する感度のパラメータを固定した状態で，提案手法の方が様々な区間長の異常に対応できることを示した．今回の比較実験では，計算時間においても提案手法が僅かに優れている結果となった．現実データを使った実験では，レビュー点数の分布変化の視覚化に成功し，得点分布の信頼性指標における有用性を示した．

第 3章では，第 2章で提案した区間分割法の大規模データへの応用として，レビューデータを用いた実験を行った．実験では，区間分割法がレビューの分析支援として十分有用であることを示した．また，今回提案した解法における解品質と計算量の関係についても深く追求した．さらに，区間分割法は高次元属性の時系列データにおいても高速で解を求めることができ，高次元の結果についても，分析支援の有効性を示した．

第 4章では，レビューサイトにおける新たなランキング手法として，提案した区間分割法の結果による区間を z-score 化し，アイテム依存の時期的な信頼を考慮した評価値を提案した．実験において，提案した basic’ (異常区間除去法)と latest (信頼区間採用法)は，basic (基本多項分布法)と比較して，それぞれ異なる性質を持つことを示した．特に latest は，レビュー投稿数やレビュー平均評点といったナイーブな情報に依存しにくい性質を持っているため，新たな評価値としての有用性が期待できる．

第 5章では，位置情報と時刻情報が正確であるデータを扱うことを前提として，第 4章で提案した z-score に対し，情報の時間的信頼を考慮することを目的として時間減衰関数を，情報の地理的信頼性を考慮することを目的として空間減衰関数を導入した．ランキング結果の分析においては，二群順位統

計量を多群に拡張し，各カテゴリを評価する方法を提案した．時空間的信頼減衰を考慮したモデルに基づくアイテムランキングは，単純な多項分布モデルに基づくアイテムランキングと比較して，地域カテゴリによる不平等性が低いことを示した．

第 6章では，第 5章で述べた時空間モデルの発展系として，観測されたデータのネットワークモデル化を試みた．モデル化実現のために，各観光スポットの人気度を導入した Lévy flight に基づいた確率モデルと，観測されたユーザ行動データからモデルのパラメータを推定する効率的な学習アルゴリズムを提案した．また，提案したモデルから得られた条件付き確率を用いて，2 種類のスポットランキング手法を提案した．レビューサイトのデータセットから生成したユーザ行動データを用いた実験では，パラメータ推定に関する詳細な検証と，提案したランキング手法とナイーブな人気度ランキングとの比較を行った．実験結果として，パラメータ推定結果は直感的に解釈が可能であることが示され，提案ランキング手法は地域カテゴリ毎の特性を見出す指標として有用であることが示された．

第 7章では，第 6章のネットワークモデルの応用として，カテゴリ間の類似度による確率ネットワーク生成法を提案した．さらに，提案確率ネットワークを分析することで，カテゴリに関する規則性や重要性を見出すことを試みた．今回提案した確率ネットワーク生成法と PageRank 値計算法は，ベースライン手法と比較して圧倒的に高速に PageRank 値を算出することができるため，多数の観測時刻におけるデータを PageRank 時系列データとして容易に扱えることを示した．また，提案した k-medoids アルゴリズムによる PageRank 時系列データのクラスタリングでは，カテゴリの役割や機能に即した出力が見られたため，カテゴリの分析手法としての有用性が期待できる．

本論文で扱った提案手法は全てバッチ処理もしくはバッチ学習の範疇に含まれているが，各手法の要所で時間計算量を抑えているため，本論文で扱ったような大規模データのデータサイズにおいてはもちろん，更に大きなデータサイズにおいても計算量自体は実用的な範囲に収まることが予想される．しかし，空間計算量については全く触れていないため，今後は，空間計算量においても実用上の問題が出ないよう，オンライン学習への対応も視野に入れて研究を進めていきたい．

謝辞

本論文は筆者が静岡県立大学大学院経営情報イノベーション研究科経営情報イノベーション専攻博士後期課程に在籍中の研究成果をまとめたものである．同専攻教授 斉藤和巳 先生には指導教員として本研究の実施の機会を与えて戴き，その遂行にあたって終始，ご指導を戴いた．同専攻教授 池田哲夫 先生並びに和歌山大学 システム工学部教授 風間一洋 先生には副指導教員としてご助言を戴くとともに本論文の細部にわたりご指導を戴いた．本論文の副査になっていただいた同専攻教授 渡邉貴之 先生にも，中間審査会において貴重なご意見を賜ることができた．本学部 斉藤研究室の各位には研究遂行にあたり日頃より有益なご討論ご助言を戴いた．ここに深謝の意を表する．

参考文献

[1] Salganik, M. J., Dodds, P. S. and Watts, D. J.: Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market, *Science*, Vol. 311, No. 5762, pp. 854–856 (2006).

[2] Ricci, F., Rokach, L., Shapira, B. and Kantor, P. B.: *Recommender Systems Handbook*, Springer-Verlag New York, Inc, New York, NY, USA (2011).

[3] Kotkov, D., Wang, S. and Veijalainen, J.: A Survey of Serendipity in Recommender Systems, *Knowledge-Based Systems*, Vol. 111, pp. 180–192 (2016).

[4] André, P., schraefel, m., Teevan, J. and Dumais, S. T.: Discovery is Never by Chance: Designing for (Un)Serendipity, *Proceedings of the 7th ACM Conference on Creativity and Cognition*, pp. 305–314 (2009).

[5] Iaquinta, L., Gemmis, M. d., Lops, P., Semeraro, G., Filannino, M. and Molino, P.: Introducing Serendipity in a Content-Based Recommender System, *Proceedings of the 8th International Conference on Hybrid Intelligent Systems*, IEEE Computer Society, pp. 168–173 (2008).

[6] Foster, A. and Ford, N.: Serendipity and Information Seeking: An Empirical Study, *Journal of Documentation*, Vol. 59, No. 3, pp. 321–340 (2003).

[7] Onuma, K., Tong, H. and Faloutsos, C.: TANGENT: A Novel, 'Surprise Me', Recommendation Algorithm, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 657–666 (2009).

[8] Zheng, Q., Chan, C. and Ip, H. H. S.: An Unexpectedness-Augmented Utility Model for Making Serendipitous Recommendation, *Advances in Data Mining: Applications and Theoretical Aspects - 15th Industrial Conference, ICDM 2015, Hamburg, Germany, July 11-24, Proceedings*, pp. 216–230 (2015).

[9] Kleinberg, J. M.: Bursty and Hierarchical Structure in Streams, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, Edmonton, Alberta, Canada*, pp. 91–101 (2002).

[10] Zhu, Y. and Shasha, D. E.: Efficient Elastic Burst Detection in Data Streams, *Proceedings of*

the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, pp. 336–345 (2003).

[11] Sun, A., Zeng, D. D. and Chen, H.: Burst Detection From Multiple Data Streams: A Network-Based Approach, *IEEE Trans. Systems, Man, and Cybernetics, Part C*, Vol. 40, No. 3, pp. 258–267 (2010).

[12] Shen, H., Wang, D., Song, C. and Barabási, A.: Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes, *Proceedings of the 28th AAAI Conference on Artificial Intelligence, July 27 -31, Québec City, Québec, Canada.*, pp. 291–297 (2014).

[13] Cormode, G., Shkapenyuk, V., Srivastava, D. and Xu, B.: Forward Decay: A Practical Time Decay Model for Streaming Systems, *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE'09)*, pp. 138–149 (2009).

[14] Papadakis, G., Niederée, C. and Nejdl, W.: Decay-Based Ranking for Social Application Content, *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST'10)*, pp. 276–281 (2010).

[15] Koren, Y.: Collaborative Filtering with Temporal Dynamics, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, pp. 447–456 (2009).

[16] Newman, M. E. J.: The Structure and Function of Complex Networks, *SIAM Review*, Vol. 45, pp. 167–256 (2003).

[17] Wasserman, S. and Faust, K.: *Social Network Analysis*, Cambridge University Press, Cambridge, UK (1994).

[18] Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, Vol. 30, pp. 107–117 (1998).

[19] Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).

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

[21] Yamagishi, Y., Okubo, S., Saito, K., Ohara, K., Kimura, M. and Motoda, H.: A Method to Divide Stream Data of Scores over Review Sites, *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, Proceedings*, pp. 913–919 (2014).

[22] Oliveira, J. G. and Barabási, A.: Human Dynamics: Darwin and Einstein Correspondence Patterns, *Nature*, Vol. 437, p. 1251 (2005).

[23] Ma, H., Zhou, D., Liu, C., Lyu, M. R. and King, I.: Recommender Systems with Social Regularization, *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*, pp. 287–296 (2011).

[24] O'Donovan, J. and Smyth, B.: Trust in Recommender Systems, *Proceedings of the 10th international conference on Intelligent user interfaces (IUI '05)*, New York, NY, USA, ACM, pp. 167–174 (2005).

[25] Goldenberg, J., Libai, B. and Muller, E.: Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth, *Marketing Letters*, Vol. 12, pp. 211–223 (2001).

[26] Kempe, D., Kleinberg, J. and Tardos, E.: Maximizing the Spread of Influence Through a Social Network, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*, pp. 137–146 (2003).

[27] Saito, K., Kimura, M., Ohara, K. and Motoda, H.: Learning Asynchronous-Time Information Diffusion Models and its Application to Behavioral Data Analysis over Social Networks, *Journal of Computer Engineering and Informatics*, Vol. 1, pp. 30–57 (2013).

[28] Sood, V. and Redner, S.: Voter Model on Heterogeneous Graphs, *Physical Review Letters*, Vol. 94, p. 178701 (2005).

[29] Even-Dar, E. and Shapria, A.: A Note on Maximizing the Spread of Influence in Social Networks, *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE'07)*, pp. 281–286 (2007).

[30] Kimura, M., Saito, K., Ohara, K. and Motoda, H.: Opinion Formation by Voter Model with Temporal Decay Dynamics, *Proceedings of 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'12)*, LNCS 7524, pp. 565–580 (2012).

[31] Moritz, H.: Geodetic Reference System 1980, *Journal of Geodesy*, Vol. 74, No. 1, pp. 128–133 (2000).

[32] Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *Ann. Math. Statist.*, Vol. 18, No. 1, pp. 50–60 (1947).

[33] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., New York, NY, USA (1995).

[34] Brockmann, D., Hufnagel, L. and Geisel, T.: The Scaling Laws of Human Travel, *Nature*,

Vol. 439, pp. 462–465 (2006).

[35] Jiang, B., Yin, J. and Zhao, S.: Characterizing the Human Mobility Pattern in a Large Street Network, *Phys. Rev. E*, Vol. 80, p. 021136 (2009).

[36] Gonzalez, M. C., Hidalgo, C. A. and Barabási, A.-L.: Understanding Individual Human Mobility Patterns, *Nature*, Vol. 453, pp. 779–782 (2008).

[37] Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J. and Chong, S.: On the Levy-Walk Nature of Human Mobility, *IEEE/ACM Trans. Netw.*, Vol. 19, No. 3, pp. 630–643 (2011).

[38] Bishop, C. M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer (2010).

[39] Seber, G. A. F. and Wild, C. J.: *Nonlinear Regression*, John Wiley & Sons (1989).

[40] Song, C., Koren, T., Wang, P. and Barabási, A.-L.: Modelling the Scaling Properties of Human Mobility, *Nature Physics*, Vol. 6, pp. 818–823 (2010).

[41] Vansteenwegen, P., Souffriau, W. and Oudheusden, D. V.: The Orienteering Problem: A Survey, *European Journal of Operational Research*, Vol. 209, pp. 1–10 (2011).

[42] Bao, J., Zheng, Y., Wilkie, D. and Mokbel, M.: Recommendations in Location-Based Social Networks: A Survey, *GeoInformatica*, Vol. 19, No. 3, pp. 525–565 (2015).

[43] Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y.: Mining Interesting Locations and Travel Sequences from GPS Trajectories, *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, New York, NY, USA, ACM, pp. 791–800 (2009).

[44] Luenberger, D. G.: *Linear and Nonlinear Programming: Second Edition*, Kluwer Academic Publishers (2003).

[45] Easley, D. and Kleinberg, J.: *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University Press, New York, NY, USA (2010).

[46] Vázquez, A.: Growing Network with Local Rules: Preferential Attachment, Clustering Hierarchy, and Degree Correlations, *Physical Review*, Vol. 67, No. 5, p. 056104 (2003).

[47] Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L.: An Analysis of Approximations for Maximizing Submodular Set Functions, *Mathematical Programming*, Vol. 14, pp. 265–294 (1978).

本論文に関する原著論文

学術論文

1. 山岸 祐己，斉藤 和巳，
   “観光レビューデータから構築した確率ネットワークによる地域分析”，
   DBSJ Journal, Vol.15, No.2, March 2017.
2. 山岸 祐己，斉藤 和巳，武藤 伸明，
   “評点時系列データの区間分割法”，
   DBSJ Journal, Vol.12, No.3, pp.1–6, February 2014.

国際会議

1. Yuki Yamagishi, Kazumi Saito, and Tetsuo Ikeda,
   “Modeling of Travel Behavior Processes from Social Media”,
   PRICAI : 14th Pacific Rim International Conference on Artificial Intelligence, pp.626–637, 2016.
2. Yuki Yamagishi, Seiya Okubo, Kazumi Saito, Kouzou Ohara, Masahiro Kimura, and Hiroshi Motoda,
   “A Method to Divide Stream Data of Scores over Review Sites”,
   PRICAI : 13th Pacific Rim International Conference on Artificial Intelligence, pp.913–919, 2014.

紀要論文

1. 山岸 祐己，斉藤 和巳，
   “確率ネットワークの時系列分析に基づくソーシャルタグの分類”，
   経営情報イノベーション研究 Vol.5, pp.1–10, October 2016.

研究会・大会

1. 山岸 祐己，斉藤 和巳，
   “観光レビューデータから構築した確率ネットワークによる地域分析”，
   第8回データ工学と情報マネジメントに関するフォーラム (DEIM), 2016.
2. 山岸 祐己，斉藤 和巳，
   “時空間情報を考慮した統計モデルに基づく観光スポットのランキング手法”，
   第14回情報科学技術フォーラム (FIT), 2015.
3. 山岸 祐己，斉藤 和巳，
   “評点時系列データの信頼区間と異常区間の検出に基づくランキング手法”，
   ネットワークが創発する知能研究会 (JWEIN’15), 2015.
4. 山岸 祐己，斉藤 和巳，湯瀬 裕昭，武藤 伸明，
   “アクセスログデータの区間分割に基づくユーザ行動分析”，
   情報処理学会第77回全国大会 (IPSJ), 2015.
5. 山岸 祐己，斉藤 和巳，武藤 伸明，
   “順位統計量に基づく楽曲動画のタグ付け傾向の分析”，
   第13回情報科学技術フォーラム (FIT), 2014.
6. 山岸 祐己，斉藤 和巳，武藤 伸明，
   “A Comparison of Video Ranking Methods Based on a Time Series Analysis of Tags Using Multi-category Order Statistics’,
   ネットワークが創発する知能研究会 (JWEIN’14), 2014.
7. 山岸 祐己，斉藤 和巳，
   “時系列データにおける分割区間探索法の性能比較”，
   情報処理学会第76回全国大会 (IPSJ), 2014.
8. 山岸 祐己，斉藤 和巳，武藤 伸明，
   “レビュー評点時系列データの変化点に基づくアイテムとユーザの分析”，
   第6回Webとデータベースに関するフォーラム (WebDB Forum), 2013.
9. 山岸 祐己，斉藤 和巳，
   “オンラインレビューサイトにおけるレビュー変化点検出法”，
   第5回Webとデータベースに関するフォーラム (WebDB Forum), 2012.

受賞など

1. “観光レビューデータから構築した確率ネットワークによる地域分析”,
   第8回データ工学と情報マネジメントに関するフォーラム 学生プレゼンテーション賞
2. “アクセスログデータの区間分割に基づくユーザ行動分析”,
   情報処理学会第77回全国大会 学生奨励賞
3. “順位統計量に基づく楽曲動画のタグ付け傾向の分析”,
   第13回情報科学技術フォーラム FIT奨励賞
4. “時系列データにおける分割区間探索法の性能比較”,
   情報処理学会第76回全国大会 学生奨励賞
5. “レビュー評点時系列データの変化点に基づくアイテムとユーザの分析”,
   第6回Webとデータベースに関するフォーラム 企業賞（サイバーエージェント賞）
6. “オンラインレビューサイトにおけるレビュー変化点検出法”,
   第5回Webとデータベースに関するフォーラム 企業賞（リクルートテクノロジーズ賞）