情報処理学会論文誌 Vol.62 No.1 346–356 (Jan. 2021)

トピックモデルを用いたテレビ視聴における トレンド分析方法の提案

坂元 哲平[1,a)] 小林 佑輔[1] 中川 慶一郎[2] 生田目 崇[3] 後藤 正幸[4]

受付日 2020年4月10日, 採録日 2020年10月6日

概要：近年，消費者の嗜好の多様化にともない，テレビ業界においても視聴者の嗜好に寄り添った魅力的な番組戦略や広告戦略を編成する必要性が増している．このような問題意識と，デジタル化によるデータの蓄積を背景にテレビ視聴データの分析事例が報告されている．一方で，従来研究では視聴履歴を用いて視聴者と番組の関係性を表現することを目的としたモデル化事例についての議論は盛んではない．そこで本研究では，両者の関係性をトピックモデルに基づくクラスタリングによってモデル化するデータ分析手法を提案する．一般に視聴者の嗜好は時間的に変化することが考えられるため，時系列を考慮したトレンドの分析を可能とするような分析法が必要である．ここで，ドラマ番組のように3カ月を1クールとして放送される番組がいっせいに変わるというテレビ特有の事象に対して，単純なクラスタリング法ではクラスタの継続性が保たれないという問題があるため，その問題に対応するトレンド分析法を提案する．さらに，得られた結果を用いた分析を直感的に行うために，サンキーダイアグラムを用いた可視化を施す．また，多様な視聴者の視聴傾向を1つのクラスタへ一意に所属させる場合と，複数クラスタへの所属を許容する場合の2つの分析法を提案し，比較を行う．最後に，提案分析法を実際のテレビ視聴データに適用し，提案法の有効性を示すとともに，結果の分析を行い視聴者のテレビ視聴行動を明らかにする．

キーワード：トピックモデル，機械学習，テレビ視聴データ

A Proposal of Trend Analysis Method for TV Viewing Data based on Topic Model

Teppei Sakamoto[1,a)] Yusuke Kobayashi[1] Keiichiro Nakagawa[2] Takashi Namatame[3] Masayuki Goto[4]

Received: April 10, 2020, Accepted: October 6, 2020

Abstract: Due to the recent diversification of consumer preferences, it has become more important to make attractive TV programs for viewers. For such purposes, it is desirable to make use of TV viewing data that have become possible to be accumulated through digitalization. Under the background, some case studies on the analysis of TV viewing data have been reported. However, few previous studies have discussed about modeling from viewing history in order to understand the relationship between viewers and TV programs. Therefore, this study proposes data analysis methods to represent the relationship between them by clustering based on topic model. Since it is considered that viewers' preferences change over time and TV programs change by seasons (such as drama programs in particular), the proposed analytical methods can adjust on the trends. Furthermore, the visualization using Sankey diagram is drawn in order to make the analysis intuitive. We also propose and compare two analytical methods for the case of viewers' viewing tendency belong to a single cluster and allowing them to belong to multiple clusters. Finally, we demonstrate a real data analysis applied the proposed method in order to show the effectiveness of our proposed method and to clarify TV viewing behaviors.

Keywords: Topic model, Machine learning, TV viewing data

[1] 株式会社エヌ・ティ・ティ・データ
NTT DATA Corporation, Koto, Tokyo 135–8672, Japan
[2] エヌ・ティ・ティ・データ先端技術株式会社
NTT DATA INTELLILINK Corporation, Chuo, Tokyo 104–0052, Japan
[3] 中央大学
Chuo University, Bunkyo, Tokyo 112–8551, Japan
[4] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan
[a)] Teppei.Sakamoto@nttdata.com

1. はじめに

近年，消費者の嗜好や余暇時間の使い方の多様化にともない，テレビ局は視聴者の嗜好に寄り添った魅力的な番組戦略を編成する必要性が増している．同様に，広告代理店や芸能事務所にとっても，視聴者の嗜好を反映した事業戦略が必要である．このような問題意識と，デジタル化によるデータの蓄積を背景に，テレビ視聴データの分析事例が報告されている．たとえば菊池ら [1] は再生方式の観点から番組視聴傾向の分析を行っている．また，番組推薦のための研究事例もあり，土屋ら [2] は視聴番組の言語情報の観点から，Xu ら [3] は 10 週間分の視聴ログを用いた視聴時間のパターンから推薦システムを検討している．また，本研究の着眼点でもある時系列に注目し，視聴パターン推移を分析した事例 [4] もある．しかし，いずれも番組を問わない視聴パターンや，個々もしくはカテゴリレベルでの番組の分析にとどまり，視聴傾向から番組の関係性をまとめながら視聴者の嗜好を把握することを目的とした分析事例は乏しい．

視聴者の嗜好に寄り添った視聴傾向を把握するためには，全体の視聴傾向を観察するのではなく，視聴者や番組を視聴傾向の類似性という観点でクラスタリングし，クラスタの特徴を分析することが必要である．ここで，視聴者は特定のジャンルの番組のみを視聴するわけではなく，複数のジャンルに対して好むことが考えられる（以後，各ジャンルを好む度合を嗜好度と呼び，嗜好度の分布を嗜好と呼ぶ）．同様に番組もいくつかの異なる側面を持つため，複数クラスタに所属できることが望ましい．このような場合に，（確率的）潜在クラスモデル [5], [6] に基づくソフトクラスタリングが有効と期待される．本研究では，確率的潜在クラスモデルのなかでも，視聴者と番組の共起を潜在クラスの下での条件付き独立と仮定し，両者を紐づけながら同時にクラスタリングすることが可能な PLSA（Probabilistic Latent Semantic Analysis）[7] を用いて，視聴傾向のクラスタリングを行う．PLSA は，文書データや購買履歴データ分析等で広く適用され，与えられたデータをよく表現するという特性から，特に分析そのものを目的とする場合に有効である．

ここで，視聴者の嗜好は時間的に変化することが一般的であり，加えてテレビ番組にはクールという概念の元で 4 半期ごとに番組の変更がある．このことから，視聴傾向についてはクラスタの時間的変化をとらえられること（＝トレンドの把握）が重要となる．一方で，PLSA を含む通常のクラスタリング手法は時間的変化を対象としていない．時間的変化の概念を導入するために，区切られた期間ごとのデータを対象に個別にクラスタリングするシンプルな方法も考えられるが，期間ごとにクラスタ全体が持つ特徴やクラスタ間の差異が完全に変わってしまうという継続性の観点で問題があり，実用的な分析を困難にする．時系列に沿った動的クラスタリング手法も提案されているが [8], [9], [10]，ドラマ番組のように，ある時点での番組の総入れ替えが生じるような問題を対象としていない．また，それらの多くの研究では興味の主眼が文書データ中の単語ベースのクラスタの推定にあり，同一文書がどのように時系列変化するかを興味の対象としていない．一方，本研究では視聴者の嗜好の変化を分析することが主眼となる．以上の理由から，既存のモデルをそのまま適用することが望ましいとはいえない．

そこで本研究では，番組が入れ替わる場合でもクラスタの継続性を維持し，視聴者の嗜好の変化をとらえられる分析方法を提案する．具体的には，視聴者が所属するクラスタが時間変化することを許容したトピックモデルの構築法と可視化手法を示す．全期間で学習したトピックモデルを用いて，各視聴者の嗜好の変化をクラスタの移動という事象で表すとともに，その推定結果をサンキーダイアグラム [12] を用いて可視化することで，直観的なトレンド分析を可能とする．最後に，提案した分析方法を実際のテレビ視聴データに適用し，提案法の有効性を示すとともに，結果の分析により新たな知見を得る．

2. 準備

2.1 対象データ

本論文では，経営科学系研究部会連合協議会主催・平成 30 年度データ解析コンペティションで提供された株式会社ビデオリサーチの VR CUBIC データ [13] を用いる．VR CUBIC データは，テレビ視聴データ，Web 接触データ，プロフィールデータとそれらに付随するマスタデータによって構成される．本研究では，テレビ視聴データおよびプロフィールデータ中の視聴者のデモグラフィック属性と各種マスタデータを利用して分析を行う．

本研究では，次の 2 点の理由によりドラマ番組の視聴データを対象とする．それは，第 1 に視聴者の嗜好性が強く出ること，第 2 に提供データのデータ取得期間（1 年間）に対し，その期間内の放送番組の変化が大きいことである．アニメ番組も同様の条件を満たすが，市場の大きさを考慮しドラマ番組を対象とした．

提供データでは，視聴者が分単位の各時点にどのチャンネルを視聴していたかが記録されている．本研究では，視聴者がどの番組を視聴したかに焦点を当てるため，視聴履歴データと番組マスタを突合し，番組の放送時間（CM 含む）の一定以上の時間を視聴した場合に視聴したものとする．それを番組 ID で統合し，行と列に視聴者と番組，要素に視聴回数（視聴話数）を持つ行列を分析の入力とする．なお，番組 ID はドラマの個々のエピソードではなく，1 つの作品やシリーズに対して 1 つ振られている．実際に分析に用いたデータのサイズ等は 4.1 節に記載する．

2.2 トピックモデル

トピックモデルと呼ばれる手法のベースは，観測されたデータの背後に観測できない潜在的な変数の存在を仮定した潜在クラスモデルである [5], [6]．潜在的な変数の仮定は，異質のデータが混ざったような現実的で複雑な問題の分析を可能とする．潜在クラスモデルのなかでも，文書と単語の共起関係の表現のために提案されたモデルがトピックモデルである．トピックモデルは，確率モデルゆえの汎用性と拡張性の高さから，文書データ以外にも顧客と商品等の購買履歴データについても応用され，購買傾向の背後にある潜在的な情報の抽出を可能としている [11]．本研究では，基本的なトピックモデルである PLSA [7] を用いる．PLSA では，共起関係にある両者を同時にクラスタリングする．これにより，単に一方を高次元スパースなベクトル間の類似度からクラスタリングする場合に比べ，他方の要素の相関関係が加味され，結果として精度の高いクラスタを構築できる．

トピックモデルを時系列に拡張したモデルも提案されている．たとえば，PLSA にベイズの概念を導入した LDA（Latent Dirichlet Allocation）[14] に時間概念を導入したモデル [8] や，さらに拡張を行い購買行動を対象としたモデル [9] がある．また，佐々木ら [10] は，SNS 上の文書データを対象に，トピックの時間変化の要素の多くを満たすモデルを提案している．これらのモデルは，文書データではある時期を境にいっせいに単語が出現しなくなることは考えづらいことから，複数期間に登場する単語を媒介としてだんだんとトピックが移り変わる様をモデル化している．いい換えれば，テレビ視聴データのような，視聴者は継続するが放送番組がある時点でいっせいに変わることに対応していない．たとえば，Blei ら [8] を本研究の対象データに適用した場合，クラスタと放送期間が対応するような学習が行われる．また，一般に文書データに対するトピックモデルでは文書ではなく単語をベースとしたトピックの推定に興味の主眼がある．たとえば佐々木ら [10] は SNS 上の投稿のクラスタリングが目的ではなく，その中に出現する単語のトレンドを追うことを主眼としている．別の視点で考えれば，投稿自体を時系列で追っていないといえ，ある時点の投稿と次の時点の投稿は別物として扱う．一方，本研究では視聴者自体は時系列で変わらず，その視聴者の嗜好がどのように推移しているかを分析することが目的である．このように，関連研究をそのまま本分析に適用することは適切ではなく，対象データ構造に合わせた分析方法が必要となる．

2.3 サンキーダイアグラム

本研究では，クラスタの時間変化の可視化を行うことで分析の可読性向上を狙う．具体的には，時間的変化のなかでも，成長や衰退等のクラスタの大きさの量的概念を表すことと，結合や分離等の他のクラスタとの関わりの概念を表すことを目的として，サンキーダイアグラム [12] を用いる．

サンキーダイアグラムは，フローダイアグラムの一種であり，工程間の流量を表現する図表である．各工程におけるノードと，異なる工程間のノードをつなぐパスで構成され，ノードとパスはともに幅（大きさ）を持つ．主にエネルギーや物資，経費等の変位を表すために用いられてきたが，近年では多分野での適用が報告されている．たとえば Malik ら [15] は，ソーシャルメディア上での話題の推移を分析するために，複数の LDA の結果をサンキーダイアグラムで可視化している．また，サンキーダイアグラム自体の可読性を向上させる研究も報告されており，たとえば David ら [16] は整数計画法を用いて，パスの重なりを考慮した図表作成を提案している．

2.4 本研究の位置づけ

以上より，本研究の位置づけは次の 3 点から説明される．(1) テレビ視聴データを対象として，視聴傾向から番組の関係性をまとめながら視聴者の嗜好を把握することを目的とした分析事例は乏しく，トピックモデルを用いたテレビ視聴データの分析には新規性がある．(2) 時系列トピックモデルの関連研究では主に文書を対象として，時系列でのトピックの推移に着目して分析をしている．この際，文書データに対するトピックモデルでは文書ではなく単語をベースとしたトピックの推定に興味の主眼がある．一方，本論文ではテレビ視聴データを扱い，分析の興味が視聴者の嗜好の推移にあるため，視聴者を嗜好でクラスタリングする．視聴者のクラスタリングを関連研究の枠組みで実施する場合，視聴者が視聴した番組のトピックから視聴者のトピック分布（嗜好度の分布）を推定する処理や，時系列上は独立として扱った視聴者を視聴者 ID 等をもとに同一視聴者として紐づける処理等が必要となる．それに対し，提案する分析方法ではシンプルに視聴者の嗜好度のパラメータを推定し，クラスタリングできる点が提案法のメリットとなる．(3) テレビ視聴データ特有の問題として，ドラマ番組等はある時点を境にいっせいに放送がなくなる．このとき既存の手法では，期間がクラスタに対応するように学習が行われてしまう．この問題に対応できる方法を提案する．

3. 提案分析方法

3.1 概要

本研究の目的は，テレビ視聴のトレンドを分析するための分析方法の提案である．具体的には，視聴者のテレビ番組視聴傾向をクラスタリングによって細分化し，視聴傾向の把握とその規模を把握する．さらに，期間ごとの視聴傾向を表現し，その推移の可視化により時間的変化を含めた

分析を可能とする方法を提案する．

提案する分析方法は次の 3 つのステップから構成される．単純に各期間で個別にクラスタリングした場合はクラスタの意味的継続性が失われてしまう．そこで，はじめに全期間の視聴履歴データで視聴傾向を PLSA で学習し，期間に依らない視聴傾向全体をモデル化する（ステップ 1）．これにより，期間を跨いでも類似する番組が含まれた全期間で意味的な継続性を持つクラスタを形成することができる．次に，全期間の視聴履歴データを各期間の視聴履歴データへと分割する．そして，ステップ 1 で推定した PLSA のパラメータを活用しながら，再度期間ごとに PLSA を学習し，視聴者を期間ごとのクラスタに割り当てる（ステップ 2）．これにより，期間ごとの視聴者の所属クラスタが明らかになるため，視聴者の人数によってクラスタの大きさを把握することができる．また，ある期間でそのクラスタに所属する番組がなければ，視聴者も割り当てられないため，クラスタの消滅を表現することができる．これはクラスタの生成についても同様である．最後に，期間の間の視聴者の遷移数をインプットにサンキーダイアグラムによる可視化を行う（ステップ 3）．これにより，クラスタの結合や分裂を表現できる．さらに，この方法はクラスタの意味を時間方向で固定するため，期間が離れていても問題なく解釈できることから，クラスタの復活（消滅後の生成）も表現可能である．

以上の 3 ステップについて次の 2 つの分析法を比較する．分析法 1 では，各期間で視聴者をクラスタに割り当てる際に最も嗜好度の強いクラスタに所属させるというシンプルな方法でクラスタリングする．これに対し分析法 2 では，視聴者は複数のクラスタに対して嗜好度を持つことができるように潜在クラスタを新たに構成し，その潜在クラスタへと割り当てる．以下で，これらの詳細を述べる．

3.2 分析法 1

まず，提案モデルで用いる変数を定義する．I 個の番組を $x_i(i = 1, \cdots, I)$，J 人の視聴者を $y_j(j = 1, \cdots, J)$，K 個の潜在的なクラスタを $z_k(k = 1, \cdots, K)$ と定義する．

ここで，視聴者 y_j が番組 x_i を視聴する事象の確率モデルは，PLSA にならって以下の式で表す．

$$P(x_i, y_j) = \sum_{k=1}^{K} P(z_k)P(x_i|z_k)P(y_j|z_k) \tag{1}$$

ただし，$P(z_k), P(x_i|z_k), P(y_j|z_k)$ にはすべて多項分布を仮定する．式 (1) は潜在変数を含むモデルであるため，EM アルゴリズム [17], [18] によってパラメータを推定する．

いま，視聴者 y_j が番組 x_i を見た回数を $n_{i,j}$ と定義すると，対数尤度関数 LL は以下の式で表される．

$$LL = \sum_{i,j} n_{i,j} \log P(x_i, y_j) \tag{2}$$

このとき，EM アルゴリズムの更新式は以下のとおりとなる．

【E ステップ】

$$P(z_k|x_i, y_j) = \frac{P(x_i, y_j, z_k)}{\sum_k P(x_i, y_j, z_k)} \tag{3}$$

【M ステップ】

$$P(z_k) = \frac{\sum_{i,j} n_{i,j} P(z_k|x_i, y_j)}{\sum_k \sum_{i,j} n_{i,j} P(z_k|x_i, y_j)} \tag{4}$$

$$P(x_i|z_k) = \frac{\sum_{j} n_{i,j} P(z_k|x_i, y_j)}{\sum_{i,j} n_{i,j} P(z_k|x_i, y_j)} \tag{5}$$

$$P(y_j|z_k) = \frac{\sum_{i} n_{i,j} P(z_k|x_i, y_j)}{\sum_{i,j} n_{i,j} P(z_k|x_i, y_j)} \tag{6}$$

次にステップ 2 として，期間ごとの視聴履歴を用いて視聴者を各クラスタに割り当てる．具体的には，全視聴履歴を期間ごとに分割した視聴履歴データを用いて再度パラメータを推定する．いま，期間を $s_l(l = 1, \cdots, L)$ に分割し，期間 s_l に視聴者 y_j が番組 x_i を視聴した回数を $n^l_{i,j}$ と定義する．このとき，期間 s_l における対数尤度 LL^l は以下の式で表される．

$$LL^l = \sum_{i,j} n^l_{i,j} \log P(x_i, y_j)^l \tag{7}$$

ここで，全期間で学習したときのパラメータを EM アルゴリズムの初期値として用いるとともに，番組に関するパラメータ $P(x_i|z_k)$ を固定する．これにより，全期間で学習したクラスタの意味合いを保持して各期間の学習ができる．期間 s_l における EM アルゴリズムの更新式は以下のとおりとなる．

【E ステップ】

$$P(z_k|x_i, y_j)^l = \frac{P(x_i, y_j, z_k)^l}{\sum_k P(x_i, y_j, z_k)^l} \tag{8}$$

【M ステップ】

$$P(z_k)^l = \frac{\sum_{i,j} n^l_{i,j} P(z_k|x_i, y_j)^l}{\sum_k \sum_{i,j} n_{i,j^l} P(z_k|x_i, y_j)^l} \tag{9}$$

$$P(y_j|z_k)^l = \frac{\sum_{i} n^l_{i,j} P(z_k|x_i, y_j)^l}{\sum_{i,j} n^l_{i,j} P(z_k|x_i, y_j)^l} \tag{10}$$

最後に，期間 s_l において視聴者 y_j に割り当てるクラスタ $c_{j,l}$ を次式で決定する．

$$c_{j,l} = \begin{cases} z_{K+1} & \sum_i n^l_{i,j} = 0 \\ \arg\max_{z_k} P(z_k|y_j)^l & \text{otherwise} \end{cases} \tag{11}$$

ただし，z_{K+1} はその期間に視聴履歴がない視聴者が所属するクラスタとして新たに定義する．また，$P(z_k|y_j)^l$ はベイズの定理により次式を用いて算出できる．

$$P(z_k|y_j)^l \propto P(z_k)^l P(y_j|z_k)^l \tag{12}$$

ここで，視聴者をあるクラスタへと一意に割り当てるの

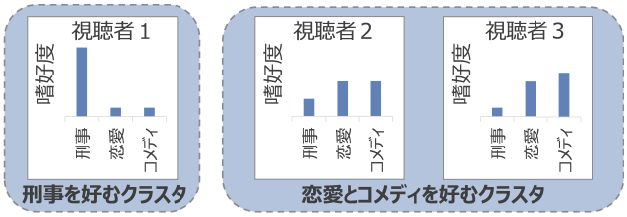

図 1 嗜好度の分布をクラスタリングするイメージ
Fig. 1 Clustering of preference distribution.

は，ステップ 3 で行うサンキーダイアグラムによる可視化のインプットとするためである．この操作により PLSA が本来持つ複数のトピックへの所属という利点は失われるが，学習の過程ではその特性を利用しているため単純にハードなクラスタリング手法を用いるより視聴傾向をとらえることが見込まれる．

ステップ 3 では，得られた $c_{j,l}$ を用いて期ごと・クラスタごとの所属人数と遷移人数を算出し，それらをインプットとしてサンキーダイアグラムで可視化を行う．

3.3 分析法 2

分析法 1 では，ステップ 2 で視聴者をある 1 つのクラスタ z_k に割り当てた．しかし，PLSA は本来複数のクラスタに確率的に所属することを許容するモデルである．いい換えれば，複数のクラスタに嗜好度を持つことができる．そこで，視聴者の嗜好を表現するために，視聴者が持つ分布を特徴量としたクラスタを新たに構築する．そのイメージを**図 1** に示す．図 1 において，分析法 1 では一意にクラスタを割り当てるため，視聴者 1 は刑事クラスタ，視聴者 2 は僅差で恋愛クラスタ，視聴者 3 はコメディクラスタへと割り当てる．分析法 2 では，視聴者 2，視聴者 3 は嗜好度の分布が類似しているため，新たに恋愛とコメディを同程度好むクラスタを構築し，そこに割り当てる．ステップに照らし合わせると，ステップ 1 とステップ 2 の間に以下の処理を行う．

いま，期間 s_l に視聴者 y_j がクラスタ z_k を選択した数を $N^l_{j,k} = \sum_{i=0}^{I} P(z_k|x_i) n^l_{i,j}$ として計算し，全期間で視聴者 y_j がクラスタ z_k を選択した回数を $N_{j,k} = \sum_{l=1}^{L} N^l_{j,k}$ と定義する．$P(z_k|x_i)$ はベイズの定理より次式を用いて算出できる．

$$P(z_k|x_i) \propto P(z_k)P(x_i|z_k) \tag{13}$$

$\boldsymbol{N}_j = (N_{j,1}, \cdots, N_{j,K})$ とすれば，$\boldsymbol{N}_j$ は全期間での視聴者 y_j の嗜好度の分布となる．ここで，視聴者の嗜好度分布 $\boldsymbol{N}_j$ を新たな潜在クラスタ $v_m (m = 1, \cdots, M)$ でクラスタリングすることを考える．$P(z_k|v_m)$ を潜在クラスタ v_m におけるクラスタ z_k の出現確率と定義し，視聴者の嗜好度分布の出現確率 $P(\boldsymbol{N}_j)$ を以下の確率モデルで表現する．

$$\begin{aligned} P(\boldsymbol{N}_j) &= \sum_{m=1}^{M} P(v_m)P(\boldsymbol{N}_j|v_m) \\ &= \sum_{m=1}^{M} P(v_m) \frac{\Gamma(\sum_k N_{j,k} + 1)}{\prod_k \Gamma(N_{j,k} + 1)} \prod_{k=1}^{K} P(z_k|v_m)^{N_{j,k}} \end{aligned} \tag{14}$$

ただし，嗜好度の分布は多項分布の同時確率に従うと考え，$P(v_m)$，$P(z_k|v_m)$ は多項分布を仮定する．すなわち嗜好度の類似度は，多項分布から発生する尤度で計算を行う．また，Γ はガンマ関数を表す．分析法 1 と同様，EM アルゴリズムによってパラメータ推定を行う．更新式の導出は仁ノ平ら [19] を参照されたい．

【E ステップ】

$$P(v_m|\boldsymbol{N}_j) = \frac{P(v_m)P(\boldsymbol{N}_j|v_m)}{\sum_{m=1}^{M} P(v_m)P(\boldsymbol{N}_j|v_m)} \tag{15}$$

【M ステップ】

$$P(v_m) = \frac{\sum_{j=1}^{J} P(v_m|\boldsymbol{N}_j)}{\sum_{m=1}^{M}\sum_{j=1}^{J} P(v_m|\boldsymbol{N}_j)} \tag{16}$$

$$P(z_k|v_m) = \frac{\sum_{j=1}^{J} P(v_m|\boldsymbol{N}_j)N_{j,k}}{\sum_{m=1}^{M}\sum_{j=1}^{J} P(v_m|\boldsymbol{N}_j)N_{j,k}} \tag{17}$$

分析法 2 のステップ 2 では，期間 s_l における視聴者 y_j をその期間の嗜好度分布 $\boldsymbol{N}^l_j = (N^l_{j,1}, \cdots, N^l_{j,K})$ からクラスタ $c_{j,l}$ に割り当てる．

$$c_{j,l} = \begin{cases} v_{M+1} & \sum_i n^l_{i,j} = 0 \\ \arg\max_{v_m} P(v_m|\boldsymbol{N}^l_j) & \text{otherwise} \end{cases} \tag{18}$$

ただし，v_{M+1} はその期間に視聴履歴がない視聴者が所属するクラスタとして新たに定義した潜在クラスタである．

ステップ 3 では，分析法 1 と同様にサンキーダイアグラムによる可視化を行う．

4. 実データ分析

本章では，提案した分析方法をテレビ視聴データに対して適用し，その結果を示す．

4.1 対象データと分析条件

対象データは 2.1 節で述べたとおり，ドラマを対象とするテレビ視聴データである．利用データの期間は 2017 年 4 月 3 日から 2018 年 3 月 31 日である．提供データでは定期的にモニタの入れ替えがあるが，本分析では年間を通してデータが取得され，ドラマ視聴履歴のある 3,247 人（$= J$）を分析対象とした．番組は，週に 1 度のレギュラー放送かつ放送時間が 25 分以上のドラマカテゴリの番組に限定した．その結果，197 番組（$= I$）が対象となった．また，視聴時の再生形式は問わず，放送時間の 50%以上の時間を視聴した場合を視聴とカウントした．期間は 1 カ月ごとに分けた（$L = 12$）．

潜在クラスの数は，複数パターンの施行の結果，解釈

表 1 エントロピーを用いた評価
Table 1 Evaluation by entropy.

	entropy
従来法	2.15
分析法 1	1.83
分析法 2	1.72

性の観点と可視化までを行う提案分析法の特徴を鑑みて $K = 10$，$M = 10$ とした．

4.2 分析方法の有効性検証

本節では，詳細な分析に入る前段として提案する分析方法の有効性を示す．定性的には，単純に期間を分けてクラスタリングを何度も行う場合と比較して，クラスタの特徴維持の観点で優位性があるといえる．また，既存の時系列を考慮したトピックモデルに対しては，番組の総入れ替えへの対応の観点で優位性があるといえる．一方で，クラスタリング問題全般の課題として，正解がないため定量的な評価は難しい．また，一般的なトピックモデルの評価指標として，トピックの予測性能を表す perplexity 等があるが，本論文の目的である視聴者のクラスタリングおよびそのクラスタの推移の把握には直接関連性があるとはいえない．

ここで，サンキーダイアグラムの視認性という観点から，エントロピーを用いて提案手法を定量的に評価する．いま，期間 s_l におけるクラスタ z_k を z_k^l と定義し，所属する視聴者数を $|z_k^l|$ と表現する．また，$P(z_k^{l+1}|z_k^l)$ をクラスタ z_k^l に所属する視聴者のうち．次の期間でクラスタ z_k^{l+1} に所属する視聴者の割合とする．すなわち，$P(z_k^{l+1}|z_k^l)$ はクラスタ z_k^l が次の期間でどの程度分裂するかを表す値である．このとき，エントロピーは次式で表される．

$$Entropy = \frac{1}{L-1}\sum_{l=1}^{L-1}\sum_{k=1}^{K+1}\frac{|z_k^l|}{J}\sum_{k'=1}^{K+1}P(z_{k'}^{l+1}|z_k^l)\log P(z_{k'}^{l+1}|z_k^l) \quad (19)$$

ただし，分析法 2 では，z_k^l を v_m^l，K を M と読み替えてエントロピーを計算する．

エントロピーが大きいことは，クラスタの分裂が多く，サンキーダイアグラムの視認性を低下させることを意味するため，本分析では小さいほうが好ましい．比較対象として，各期間で独立にクラスタ数 K の PLSA を実施した場合（従来法）をベースラインとして想定する．**表 1** に，初期値をランダムに設定し，10 回の繰り返し実験の平均結果を示す．

表 1 より，提案する分析方法がベースラインに比べてエントロピーが低下していることが分かる．このことから，全期間でモデルを学習後に，各期間のデータを用いて視聴者をクラスタに割り当てる提案法により，クラスタの特徴維持という期待した振舞いをし，結果として視認性が向上していることが示された．分析法 1 と分析法 2 を比較すると，分析法 2 が小さい値をとっている．すなわち，最も嗜好度の強いクラスタの変化に比べて，嗜好度の分布の変化は小さいことを意味している．

4.3 分析法 1 の結果

はじめにステップ 1 の結果を示す．PLSA は番組と視聴者を同時にクラスタリングするため，両側面から分析が可能である．**表 2** に各クラスタでの出現確率が上位の 5 番組を列挙する．番組名の末尾は放送時期を示している（Q はクールを表し，各クールの期間は次のとおり．Q1：2017 年 4〜6 月，Q2：2017 年 7〜9 月，Q3：2017 年 10〜12 月，Q4：2017 年 1〜3 月）．**表 3** に各クラスタに所属する視聴者の属性の期待値を示す．

表 2 より，PLSA を適用することで，各クラスタがそれぞれ異なる観点のもと類似する番組がまとめられていることが分かる．その解釈を表の先頭に記載した．たとえば，クラスタ z_1 は大河ドラマを好むクラスタと解釈できる．クラスタ z_2, z_5, z_8 はどれも平日の 21 時から 23 時ごろに放送されている一般的なドラマであるが，クラスタ z_2 はすべて女性が主演のドラマ，クラスタ z_5 は重厚な恋愛やヒューマン系に重きを置いたドラマ，クラスタ z_8 はすべて男性主演のドラマといったようにクラスタリングされている．また，クラスタ z_6, z_9 はどちらも刑事ドラマを好むクラスタであるが，クラスタ z_6 はテレビ朝日放送のドラマ，クラスタ z_9 はそれ以外というようにクラスタ分けされている．また，すべてのクラスタにおいて，クラスタを特徴づける番組が同一クールの番組のみではないことも分かる．

表 3 より，視聴者の側面では，いくつかのクラスタで特徴が出ている．はじめに平均年齢を見ると，全体の平均が 43.8 歳に対し，どのクラスタの平均年齢も 40 代と大きな差がない．次に性別を見ると，全体の男性比率 0.51 に対し，z_2 では男性比率が低く，女性主演のドラマ好む男性が少ない傾向にあることが分かる．また，z_4 では男性比率が高く，深夜ドラマや海外ドラマを男性が好む傾向にあることが分かる．ただし，いずれも 0.1 程度の差であることから，極端に男性（もしくは女性）のみが好むクラスタは存在しないことが分かる．未婚比率についても，全体で 0.29 に対し，z_4 では 0.39 と高い比率になっている．このことから，深夜ドラマや海外ドラマは未婚の男性が好む傾向にあると特徴づけられる．このように，視聴者の属性と番組を紐づけることができるクラスがある．一方で，多くのクラスタでは視聴者の属性の極端な差は見られないことも分かる．

次に，ステップ 2 および 3 の結果を**図 2** に示す．図 2 は，左から右に時間が推移し，1 カ月ごとのクラスタの大きさと視聴者のクラスタ間の移行を表している．▼はクールの切り替えを示す．また，上から順にステップ 1 で得られたクラスタを表し，一番下のクラスタは視聴なしクラス

表 2 年間の視聴履歴のクラスタリング結果（番組特徴：$P(x_i|z_k)$ の上位 5 番組）
Table 2 Clustering of viewing history throughout the year (Characteristic of TV program: $P(x_i|z_k)$ Top 5).

z_1 大河ドラマ	z_2 一般（女性主演）	z_3 夜中	z_4 深夜・海外	z_5 一般（恋愛・ヒューマン）
おんな城主直虎（Q1, Q2, Q3）	火曜ドラマ・あなたのことはそれほど（Q1）	木曜ドラマ・恋がヘタでも生きてます（Q1）	ドラマ 24・孤独のグルメ Season6（Q1）	金曜ドラマ・リバース（Q1）
西郷どん（Q4）	過保護のカホコ（Q2）	木曜ドラマ F・リピート・運命を変える（Q4）	コードネームミラージュ（Q1, Q2）	日曜劇場・ごめん、愛してる（Q2）
日曜劇場・陸王（Q3）	火曜ドラマ・カンナさーん！（Q2）	木曜ドラマ・脳にスマホが埋められた！（Q2）	GRIMM/グリムシーズン 1（Q2, Q3）	愛してたって、秘密はある。（Q2）
土曜ドラマ・植木等とのぼせもん（Q2）	奥様は、取り扱い注意（Q3）	木曜ドラマ F・ブラックリベンジ（Q3）	ドラマ 24・下北沢ダイハード（Q2）	僕たちがやりました（Q2）
日曜劇場・小さな巨人（Q1）	木曜劇場・人は見た目が 100 パーセント（Q1）	金曜ナイトドラマ・女囚セブン（Q1）	木ドラ 25・さぼリーマン甘太朗（Q2）	anone（Q4）
z_6 テレビ朝日刑事	z_7 NHK	z_8 一般（男性主演）	z_9 一般（刑事）	z_{10} ヒット作
相棒（Q3, Q4）	ドラマ 10・この声をきみに（Q3）	ボク、運命の人です。（Q1）	CRISIS 公安機動捜査隊特捜班（Q1）	コード・ブルー・ドクターヘリ緊急救命（Q2）
木曜ミステリー・科捜研の女（Q3, Q4）	ドラマ 10・ツバキ文具店・鎌倉代書屋物語（Q1）	もみ消して冬・わが家の問題なかったことに（Q4）	日曜劇場・小さな巨人（Q1）	金曜ドラマ・コウノドリ（Q3）
木曜ドラマ・ドクター X・外科医 大門未知子（Q3）	土曜ドラマ・4 号警備（Q1）	先に生まれただけの僕（Q3）	日曜劇場・99.9・刑事専門弁護士（Q4）	日曜劇場・陸王（Q3）
刑事 7 人（Q2）	大河ファンタジー・精霊の守り人・最終章（Q3）	ウチの夫は仕事ができない（Q2）	木曜劇場・刑事ゆがみ（Q3）	木曜ドラマ・ドクター X・外科医 大門未知子（Q3）
木曜ミステリー・遺留捜査（Q2）	ドラマ 10・ブランケット・キャッツ（Q2）	フランケンシュタインの恋（Q1）	貴族探偵（Q1）	金曜ドラマ・アンナチュラル（Q4）

表 3 年間の視聴履歴のクラスタリング結果（視聴者特徴）
Table 3 Clustering of viewing history throughout the year (Characteristic of viewer).

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	全データ
平均年齢	48.4	41.4	41.7	43.4	41.6	47.4	47.6	40.3	43.7	42.0	43.8
男性比率	0.58	0.40	0.53	0.65	0.44	0.52	0.56	0.47	0.50	0.48	0.51
未婚比率	0.28	0.25	0.35	0.39	0.26	0.27	0.35	0.31	0.26	0.24	0.29

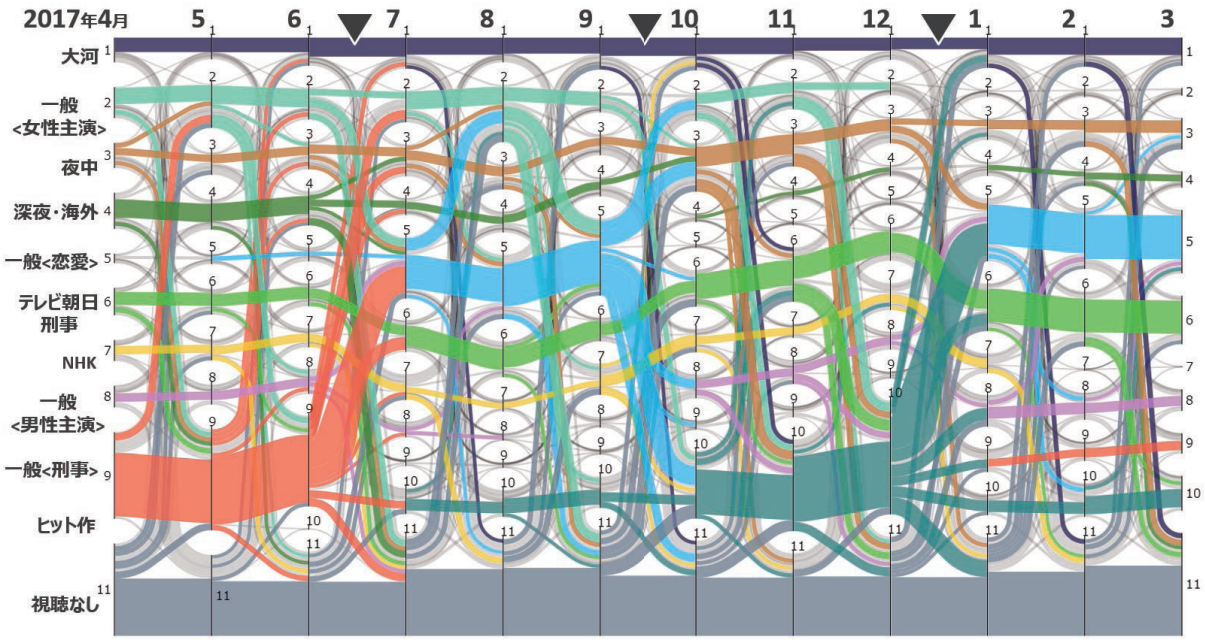

図 2 分析法 1 によるトレンドの可視化結果
Fig. 2 Trend visualization by method1.

タを表す．図中の太字の数字は年月を表し，細字の数字はクラスタ番号を表す．サンキーダイアグラムの視認性を高めるため，移行人数が全視聴者の 1%（32 人）以上のときに色を付けて描画した．

図 2 より，クラスタごとに大きさが異なることが分かる．また，クラスタ z_1（大河ドラマ）やクラスタ z_3（深夜ドラマ），クラスタ z_7（NHK ドラマ）のように大きさが時間的に安定しているクラスタもある一方で，クラスタ z_5（恋愛系）やクラスタ z_9（刑事系）のように不安定なクラスタもある．次に視聴者の移行に注目すると，全体的に他のクラスタへの移行が多く観測されることが分かる．これは，多くの視聴者の最も好む嗜好が移り変わりやすいということを示唆している．また，クールの切り替え時に最も移行数が大きいことが分かる．クール内でも，1 カ月目から 2 カ月目で視聴者が定着するクラスタと，2 カ月目から 3 カ月目で定着するクラスタが存在することが分かる．特にパスが太くなっている部分については，後者の傾向がある．これは，多くの視聴者に見られているドラマは途中からの参入者も継続視聴する傾向にあり，一方で視聴者の少ないドラマは後半で見切られてしまう傾向にあることを示唆している．パスの太い箇所に注目すると，クラスタ z_9（Q1）→クラスタ z_5（Q2）→クラスタ z_{10}（Q3）→クラスタ z_5（Q4）と見てとれる．Q1 から Q3 についてはすべて “日曜劇場” が関与しており，その枠やテイストを好む視聴者がいたと考えられる．一方で Q4 での “日曜劇場” はクラスタ z_9 であるが，Q4 の “日曜劇場” ドラマは Q1～Q3 のものと比べてライトな内容であり，それまで見ていた視聴者の嗜好と一致しなかった可能性が考えられる．

個別のクラスタに注目する．たとえば，クラスタ z_1（大河ドラマ）は年間を通して大きさが安定している．これは放送番組が Q1～Q3 で変わらないことに加え，Q4 で放送番組が変わっても嗜好が変わりにくいことを示している．クラスタ z_9（ヒット作ドラマ）は，初期にはクラスタの大きさがきわめて小さいが，Q2 ごろから生成されたクラスタだといえる．また，クラスタ z_5（恋愛ドラマ）は Q3 で一度消滅し，Q4 で復活している様が分かる．最後に，視聴なしクラスに注目すると，Q2 は比較的所属人数が多い．すなわちドラマがやや不作だった可能性が示唆される．

以上のように，提案分析方法を用いることで時間的変化を考慮した視聴傾向の分析が可能となる．

4.4 分析法 2 の結果

分析法 2 では，初めに視聴者の嗜好度分布のクラスタリングを行う．その結果を**図 3** に示す．図 3 は，縦軸にクラスタ z_k（の解釈），横軸にクラスタ v_m を取り，$P(z_k|v_m)$ の値を表示し，値の大小に応じて濃淡を付けている．多項分布であるため，横軸方向の和が 1 となる．得られる解釈を図 3 の右側に記載する．

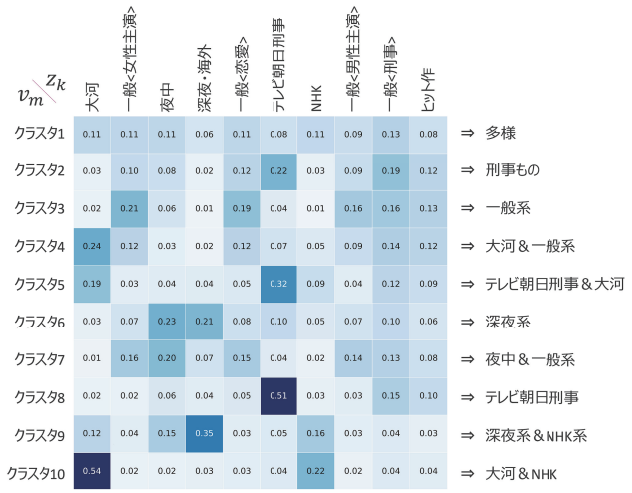

図 3 嗜好度分布のクラスタリング結果
Fig. 3 Result of clustered preference distribution.

図 3 より，いくつかの特徴的なパターンのクラスタが構築されていることが分かる．それは，特定のクラスタ z_k を好むクラスタ v_8, v_{10}，いくつかのクラスタ z_k を好むクラスタ v_5, v_9，多くのクラスタ z_k を視聴するそのほかのクラスタ v_m というようなパターンである．すなわち，クラスタ z_5（大河）やクラスタ z_8（テレビ朝日の刑事ドラマ）はそれしか見ないような視聴者が多く存在することを表している．これは，図 2 とも整合性がとれている．そのほかにも，クラスタ v_6 は夜の遅い時間にドラマを視聴するクラスタと解釈できる．また，クラスタ v_2 はテレビ局にかかわらず刑事ドラマを好むようなクラスと解釈できる．ただし，全体的に視聴者は特定のクラスタ z_k や特定の複数クラスタ z_k のみを好むわけではなく，多くのドラマを横断して視聴していることが分かる．

図 4 に嗜好度分布のクラスタリングの結果を用いて，図 2 と同様の手順で描画したグラフを示す．図 4 より，図 2 と比較すると全体的にクラスタの大きさが安定していることが分かる．すなわち，放送番組や時期によらずクラスタの嗜好度分布の変化は少ないといえる．クールの切り替え時に注目しても，図 2 に比べると遷移が少ないことが分かる．個別のクラスタに注目すると，クラスタ v_3（一般系）が最も所属人数が大きいことが分かる．すなわち，民放の夜に放送されている番組をつまみ食いするように視聴する視聴者が多いといえる．

以上の分析法 1，2 の結果から，最も嗜好度が強いジャンルは変化するものの，視聴者の嗜好度の分布自体は放送番組によって大きく変わらないことが分かる．また，どちらの観点においても，クール切り替え時には，クール内よりも嗜好が変わりやすいことが分かる．

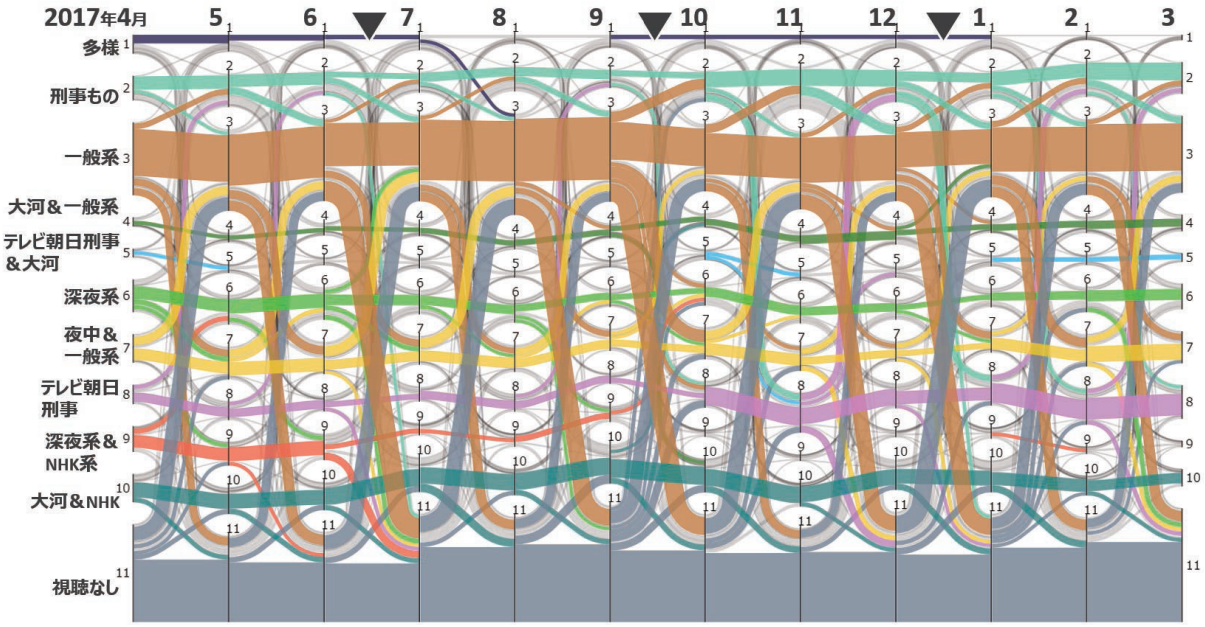

図 4 分析法 2 によるトレンドの可視化結果
Fig. 4 Trend visualization by method2.

5. 考察

5.1 提案分析方法

本論文では，2 つの分析法を提案した．分析法 1 では視聴者を最も好む番組クラスタに割り当てたことに対し，分析法 2 では最も嗜好度の分布が類似するクラスタに割り当てる．どちらの方法がより有益であるかは，分析者や分析目的によって異なるため，一概に決めることはできない．分析法 1 はシンプルにトレンドを表現しているため，視聴者の一番強い嗜好の把握が目的であれば有益な図となる．分析法 2 は，柔軟に視聴者の嗜好をとらえることができるメリットがある一方で，解釈が 2 段階になることが直観的な理解を難しくする要因になりうる．また，4.2 節の結果では分析法 1 よりも分析法 2 のエントロピーが小さいことは，分析法 2 が優れていることを示すわけではないことに注意する必要がある．ここでは，エントロピーがある程度小さいことは重要だが，それを満たせば，エントロピーの値が低ければよいというわけではなく，上述の分析目的と合わせて分析法を選択することが重要である．

なお，分析法 2 では段階的にクラスタリングをしているため，階層構造を仮定した Li ら [20] のモデルの適用が考えられる．しかし，本研究ではトピックの階層構造は仮定しておらず，あくまで視聴者が複数トピックへの嗜好を持つ特徴を保持しながら一意のクラスへと割り当てするために 2 度クラスタリングをしているという点で構造が異なっている．

また，分析を目的としたクラスタリング問題全般の課題として，クラスタ数の決定方法には議論の余地がある．本研究では可視化までを実施する特徴から，現実的に分析が可能なクラスタ数を選択することが好ましい．適切な評価指標を定め，クラスタ数を決定する方法は今後の課題としたい．

5.2 対象データの考察と他データへの適用

4 節の分析では，ドラマに限定した視聴履歴から，視聴傾向によって異なる観点でグループ化できることが分かった．加えて，クールの切り替え時に明確に嗜好の変化があることが確認できた．これらの結果を次のように活用することが考えられる．番組編成戦略では，遷移しやすいクラスの番組を絶えず提供することで視聴者を獲得することができる．また，広告戦略ではクールが変わっても視聴者を追って広告を打つことができるようになる．

結果の記載は割愛するが，ドラマに限定せずすべての番組カテゴリを対象に分析法 1 を適用したところ，興味深い結果は得られなかった．これは，ステップ 1 で粒度の大きなクラスタが構成されてしまい，1 年間というデータ取得期間に対して，大きな粒度では嗜好があまり変わらないことを示唆している．しかし，より長期間のデータがあれば，年単位でのトレンド等を分析できる可能性がある．

一方，テレビ視聴データの特性として，視聴者の嗜好が必ずしもデータに反映されているとは限らない点があげられる．テレビという特性上，生活の一部としてとりあえずテレビをつけるという行動と，放送されている番組しか視聴できないこと（番組変更，選択肢の少なさ）に起因すると考えられる．後者によって，選択肢が少ないために番組視聴 1 回がモデルに与える影響が大きくなり，結果として視聴者が他クラスへ遷移しやすくなると考えられる．その状況は可視化によっても確認できている．

他データへの適用として，EC サイトの購買履歴データや動画サイトの閲覧履歴データ等が考えられる．それらのデータはユーザの嗜好がテレビに比べると強く，また選択肢も多いため，提案法により分かりやすい結果が得られると推測される．

5.3 本論文の目的との対応

本論文の目的は，クラスタの特徴を維持しながら時間変化（成長・衰退・生成・消滅・復活・結合・分離）するトレンドを分析することである．クラスタの特徴維持は，初めに全期間でクラスタリングを行うことによって対応している．成長や減衰はクラスタの大きさから分析が可能である．生成・消滅は，厳密ではないもののクラスタの大きさで把握可能である．厳密化のためには閾値で切るほか，正則化のテクニックの導入も考えられる．クラスタの特徴維持と生成・消滅を合わせて，復活も表現が可能である．また，視聴者をクラスタリングすることによって，視聴者の嗜好の変化をとらえられ，さらにサンキーダイアグラムによる可視化によって，クラスタの結合や分離を視覚的に理解することができる．加えて，視聴なしのクラスタを設定することにより，視聴者の新規参入や離反についても分析可能となっている．

ただし，あくまでデータから推定したパラメータのみを用いた分析であることに注意し，真に視聴者の嗜好の変化をとらえられたとはいい切れない．この点は，アンケートデータ等から得られるサイコグラフィック情報を用いた統合的な分析を実施することが好ましく，今後の課題としたい．

6. まとめと今後の課題

本研究では，テレビ視聴におけるトレンドを分析することを目的とし，視聴者が所属する潜在クラスが時間的に変化する状況を表現可能なモデルを提案した．提案法では，ドラマ番組のように期間ごとに番組が総入れ替えされてしまうような状況でも分析が可能である．また，視聴者の各期間で所属する潜在クラスを推定する方法として，視聴者の嗜好を単純に考える方法と，詳細に考える方法の 2 つを提案した．さらに，提案した分析方法を実データに対して適用し，トレンドの分析を行った．その結果として，視聴傾向によるドラマ番組が異なる観点からグルーピングされることや，視聴者の嗜好の時間的変化が確認され，いくつかの知見を得ることができた．以上のことから，潜在的な嗜好の時間変化をとらえることが可能な分析手法を構築できたという点で，提案法の有効性が示されたと考える．本研究においてテレビ視聴履歴の実データを用いて得られた分析結果は，今後，実務に活用していくことが期待される．

今後の課題として，得られた結果の要因を番組の特徴をさらに加味して分析する必要がある．たとえば，番組起点では出演者等の分析を加えることや，視聴者起点では，サイコグラフィック情報等を加えた分析が必要であり，それらを踏まえた提案法の拡張が望まれる．また，本研究に対して適切な評価指標を定め，クラスタ数の決定方法を考えることも必要である．クラスタリング問題に対する評価指標は提案されているが，宇野 [21] も指摘するように，実用的な分析と指標を対応付けすることが必要である．4.2 節で実施したエントロピーによる評価は 1 つの指針となるが，本分析の目的であるトレンドの把握の程度についての直接的な指標とはなっていないことに注意し，新たな指標を考案する必要がある．

参考文献

[1] 菊池匡晃，坪井創吾，中田康太：大規模テレビ視聴データによる番組視聴分析，情報処理学会デジタルプラクティス，Vol.7, No.4, pp.352–360 (2010).

[2] 土屋誠司，佐竹純二，近間正樹，上田博唯，大倉計美，蚊野浩，安田昌司：TV 番組推薦システムの構築とその有用性の検証，情報処理学会研究報告ヒューマンコンピュータインタラクション，Vol.117, pp.95–102 (2006).

[3] Xu, M., Berkovsky, S., Koprinska, I., Ardon, S. and Yacef, K.: Time Dependency in TV Viewer Clustering, *Proc. 20th International Conference on User Modeling, Adaptation, and Personalization, UMAP 2012*, 10 pages (2012).

[4] 水岡良彰，中田康太，折原良平：大規模テレビ視聴データによる視聴パターン推移の分析，人工知能学会全国大会（第 32 回），pp.1P203–1P203 (2018).

[5] Collins, L.M. and Lanza, S.T.: Latent Class and Latent Transition Analysis: With Applications in the Social, *Behavioral, and Health Sciences*, John Wiley & Sons (2010).

[6] 岩田具治：確率的潜在変数モデルに基づくデータマイニング，オペレーションズ・リサーチ，Vol.64, No.5, pp.272–277 (2018).

[7] Hofmann, T.: Probabilistic Latent Semantic Analysis, *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, pp.286–296 (1999).

[8] Blei, D.M. and Lafferty, J.D.: Dynamic Topic Models, *Proc. 23rd International Conference on Machine Learning*, pp.113–120 (2006).

[9] Iwata, T., Watanabe, S., Yamada, T. and Ueda, N.: Topic Tracking Model for Analyzing Consumer Purchase Behavior, *21st International Joint Conference on Artificial Intelligence*, pp.11–17 (2009).

[10] 佐々木謙太朗，吉川大弘，古橋　武：複数のトピックの時間的依存関係を考慮した時系列混合モデル，人工知能学会論文誌，Vol.30, pp.466–472 (2015).

[11] 佐藤　圭：マーケティング研究におけるトピックモデルの適用に関する一考察，経営研究，Vol.68, No.3, pp.125–148 (2017).

[12] Schmidt, M.: The Sankey Diagram in Energy and Material Flow Management: Part I: History, *Journal of Industrial Ecology*, Vol.12, No.1, pp.82–94 (2008).

[13] 株式会社ビデオリサーチ：VR CUBIC，入手先 ⟨https://www.videor.co.jp/service/communication/vrcubic.html⟩（参照 2020-03-15）.

[14] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[15] Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C. and Shneiderman, B.: TopicFlow: Visualizing Topic Alignment of Twitter Data over Time, *Proc. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp.720–726 (2013).
[16] Zarate, D.C., Bodic, P.L., Dwyer, T., Gange, G. and Stuckey, P.: Optimal Sankey Diagrams via Integer Programming, *2018 IEEE Pacific Visualization Symposium* (*PacificVis*), pp.135–139 (2018).
[17] Dempster, A.P. Laird, N.M. and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B* (*Methodological*), Vol.39, No.1, pp.1–38 (1977).
[18] McLachlan G.J. and Krishnan, T.: *The EM Algorithm and Extensions*, John Wiley & Sons (2007).
[19] Ninohira, M., Yamashita, H. and Goto, M.: Customer Clustering Based on a Latent Class Model Representing Preferences for Item Seasonality, *Journal of Japan Industrial Management Association*, Vol.69, No.4E, pp.195–206 (2019).
[20] Li, W. and McCallum, A.: Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, *Proc. 23rd International Conference on Machine Learning*, pp.577–584 (2006).
[21] 宇野毅明：見過ごされてきた現場の問題—真に有益なクラスタリングを目指して— クラスタリングの問題点，オペレーションズ・リサーチ，Vol.64, No.5, pp.278–282 (2019).

坂元 哲平

2018 年早稲田大学大学院創造理工学研究科修士課程修了．同年株式会社エヌ・ティ・ティ・データ入社．データサイエンス分野の研究開発およびビジネス適用に従事．

小林 佑輔

2009 年千葉大学大学院融合科学研究科修士課程修了．同年株式会社エヌ・ティ・ティ・データ入社．ビジネスアナリティクス，AI ソリューションの研究開発およびビジネス適用に従事．

中川 慶一郎

1992 年早稲田大学大学院理工学研究科修士課程修了．同年株式会社エヌ・ティ・ティ・データ入社．2000 年早稲田大学大学院理工学研究科博士課程満期退学．2012 年株式会社 NTT データ数理システム取締役．2019 年エヌ・ティ・ティ・データ先端技術株式会社執行役員．AI ソリューション提供およびビッグデータ基盤構築のビジネスに従事．博士（工学）．日本オペレーションズ・リサーチ学会フェロー，日本経営工学会，日本経営システム学会各会員．

生田目 崇

1999 年東京理科大学大学院工学研究科博士後期課程修了．博士（工学）．同年東京理科大学助手．2002 年専修大学商学部専任講師．2013 年より中央大学理工学部経営システム工学科教授．マーケティング，経営科学の研究に従事．著書に『マーケティングのための統計解析』，オーム社 (2017)，『マーケティング・エンジニアリング入門』，有斐閣 (2017) 等．INFORMS，日本オペレーションズ・リサーチ学会，経営情報学会等各会員．

後藤 正幸 （正会員）

1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学大学院理工学研究科博士課程修了．博士（工学）．1997 年同大学理工学部助手．2000 年東京大学大学院工学系研究科助手．2002 年武蔵工業大学環境情報学部助教授．2008 年早稲田大創造理工学部経営システム工学科准教授．2011 年同大学教授．情報数理応用とデータサイエンス，ならびにビジネスアナリティクスの研究に従事．著書に，『入門パターン認識と機械学習』，コロナ社 (2014)，『ビジネス統計～統計基礎とエクセル分析』，オデッセイコミュニケーションズ (2015) 等．IEEE，INFORMS，電子情報通信学会，人工知能学会，日本経営工学会，経営情報学会等各会員．