論文・事例研究

消費者の複数メディア消費行動の統合的分析モデル

里村　卓也

1. はじめに

近年は消費者による複数のメディア消費が広がっている．総務省情報通信政策研究所 [1] による 2017 年の調査ではテレビ（リアルタイム）の平均視聴時間は平日 159.4 分，休日 214.0 分であるが，インターネットの閲覧時間も平日 100.4 分，休日 123.0 分であり，消費者はテレビ視聴だけでなくインターネット閲覧へも多くの時間を消費していることがわかる．このため，消費者のメディア消費行動を考える際には，複数メディアの消費を統合的に分析する必要がある．さらに消費者によって番組や Web サイトの消費状況は大きく異なるため，消費者個人内での各メディアの消費の特徴を知るだけでなく，メディア間での消費内容の関連性も同時に考える必要がある．

そこで本研究では，個人別の視聴番組と閲覧 Web サイトを同時に分析することで，統合的な消費者インサイトを得る方法を開発する．このとき TV 番組や Web サイトはアイテム数が多いため，統計的潜在意味解析で開発された統計的手法であるトピックモデリングによる情報の縮約を行う．マーケティング分野でもトピックモデリングによる消費者行動分析（例えば Büschken and Allenby [2], Jacobs et al. [3], Trusov et al. [4], Ansari et al. [5], 里村 [6]）が試みられており，本研究でもトピックモデルを利用して消費者の複数のメディア消費行動の統合的な分析を行う．

2. モデル

2.1 ジョイント LDA モデルについて

本研究ではトピックモデルのひとつであるジョイント LDA モデル (Blei and Jordan [7], Mimno et al. [8], Iwata et al. [9], Pyo et al. [10] 里村 [6]) を利用する．ジョイント LDA モデルは 1 つの個体についての複数のデータを統合するために Latent Dirichlet Allocation (LDA) モデル (Blei et al. [11]) から発展したものである．ジョイント LDA モデルは，言語解析 (Mimno et al. [8])，ファッション・コーディネイト (Iwata et al. [9])，ソーシャル TV(Pyo et al. [10])，顧客データ (里村 [6]) などの複数データの同時分析において利用されている．

本研究では TV 番組と Web サイトを共通して説明できるトピックを得るためにジョイント LDA モデルを利用する．ジョイント LDA モデルにより TV 番組と Web サイトに共通した潜在的トピックが得られ，これをもとに消費者の複数メディア消費行動の統合的分析を行うことが可能となる．さらに TV 番組間や Web サイト間の潜在的な共起関係から，TV 番組と Web サイトそれぞれについての潜在的利用者を抽出する．

2.2 既存分析手法（縮約データのクラスター分析）と LDA モデルの違い

変数数が大量にある多変量データをグループ分けする手法として，次元縮約（因子分析，主成分分析，多次元尺度構成法など）の結果（縮約データ）をクラスター分析により分割する方法がある．このような手法は Tandem Clustering (Arabie and Hubert [12]) や縮約データのクラスター分析 (岡太と守口 [13]) と呼ばれている．複数の多変量解析の手法を組み合わせて利用することは分析者にとっては手軽であるが留意すべき点もある．岡太と守口 [13] の指摘によると，縮約データのクラスター分析では，縮約前の完全データがもつ情報のうちの一部が縮約により失われ，このような完全ではない情報をもとにクラスター分析が行われる．一方，完全データを利用した場合には縮約データでは失われてしまった情報によってもクラスターが作られることがある．このように縮約データのクラスター分析では，完全データによるクラスター分析とは異なる結果が得られることがありうる．また黒木と山下 [14] によると，第 1 ステップ（次元縮約）と第 2 ステップ

さとむら　たくや
慶應義塾大学商学部
satomura@fbc.keio.ac.jp
受付 19.7.13　採択 19.10.30

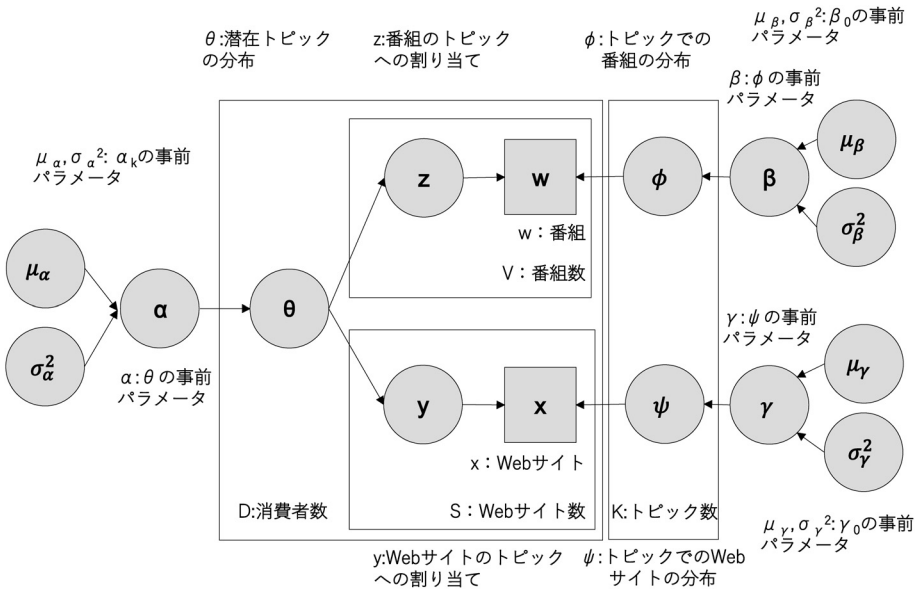

図 1　ジョイント LDA モデルのグラフィカル表現

（クラスター分析）では分析の目的関数が異なるため，第 1 ステップで得られた結果が第 2 ステップを実施するのに有用な情報となっているとは限らない．

LDA モデルでは，次元縮約とグループ化において完全データをそのまま利用し，さらに次元縮約とグループ化で同じ目的関数を利用して分析を行うことができる．

2.3　モデルの定式化

本研究では消費者は視聴番組と閲覧 Web サイトについて共通した潜在的トピックを持っているとする．消費者は，番組視聴機会毎（例えば 1 日に 5 つの番組を視聴するのであれば 5 回の視聴機会毎）に，トピックを選び，そのトピックに基づき，視聴する番組を決定するものとする．各消費者は複数のトピックを確率的にもっており，トピックの分布は視聴者毎に異なるとする．さらに各番組視聴の確率はトピックにより異なるものとする．Web 閲覧についても同様に考える．

佐藤 [15] によると，トピックモデルを用いた統計的潜在意味解析では，複数の単語の共起性によって創発される情報を「潜在的意味」と考える．なお，この共起性はデータに実際に現れる顕在的共起だけでなく，データ上には現れない隠れた共起性である潜在的共起性も考慮している．そして「潜在的意味のカテゴリー」のことをトピックと呼ぶ．本研究で用いるジョイント LDA モデルでは，視聴番組の共起性と閲覧 Web サイトの共起性によって創発される情報は両者に共通した「潜在的意味」をもっていると考える．

トピックモデルにおいて，トピックが何を表しているのかは，適用するデータとその文脈によって変わってくる．佐藤 [15] によると潜在的意味解析の分野ではトピックは「潜在的意味のカテゴリー」を表していると考える．購買商品にトピックモデルを適用した Jacobs et al. [3] では，トピックは商品購買への「モチベーション」と考えた．また，購買商品とアンケート調査のデータに適用した里村 [6] では，トピックは「潜在的ライフスタイルのカテゴリー」と考えた．本研究では，Jacobs et al. [3] のように消費者による選択行動データを利用するが，消費の対象が商品ではなくメディアであるので，トピックはメディア消費への「モチベーション」と考えることができる．

続いて，ジョイント LDA モデルを定式化する．図 1 は本モデルのグラフィカル表現である．

消費者 $d(=1,\ldots,D)$ がトピック $k(=1,\ldots,K)$ に所属する確率（トピック k の構成比率）を θ_{dk} とする．$\theta_d=(\theta_{d1},\ldots,\theta_{dK})'$ とし，θ_d の事前分布をパラメータ α のディリクレ分布とする．$\alpha_k(>0)$ は α の k 番目の要素であり，$\alpha=(\alpha_1,\ldots,\alpha_K)'$ とする．また α_k の事前分布をパラメータ $\mu_\alpha,\sigma_\alpha^2$ に従う対数正規分布とする．

$$\theta_d \sim \mathrm{Dirichlet}(\alpha)$$
$$\alpha_k \sim \mathrm{LogNormal}(\mu_\alpha,\sigma_\alpha^2)$$

このように θ_d の事前分布のパラメータ α は非対称に設定されており，各 α_k はデータから推定する．

次に視聴番組に関する定式化を行う．メディア消費へ

のモチベーションであるトピックによって各番組の視聴のされやすさが異なるとする．トピック $k(=1,\ldots,K)$ における番組 $v(=1,\ldots,V)$ の出現確率を ϕ_{kv} とする．$\phi_{kv}=(\phi_{k1},\ldots,\phi_{kV})'$ とし，ϕ_k の事前分布をパラメータ β のディリクレ分布とする．β は共通の要素 $\beta_0(>0)$ からなるサイズ V のベクトルとする．

$$\phi_k\sim\text{Dirichlet}(\beta)$$
$$\beta_0\sim\text{LogNormal}(\mu_\beta,\sigma_\beta^2)$$

このように，ディリクレ分布のパラメータは対称である．また β_0 の事前分布をパラメータ μ_β,σ_β^2 に従う対数正規分布とする．β_0 はデータから推定する．

消費者 d の番組 v の期間中の総視聴回数を N_{dv} とする．すると $N_d=\sum_{v=1}^{V}N_{dv}$ は消費者 d の期間中の全番組の総視聴回数となる．

消費者 d の $n(=1,\ldots,N_d)$ 番目の番組視聴機会におけるトピックを z_{dn} とする．z_{dn} は離散値をとる潜在変数であり，パラメータ θ_d の多項分布に従うとする．また消費者 d の n 番目の視聴機会における視聴番組を w_{dn} とする．w_{dn} はパラメータ $\phi_{z_{dn}}$ の多項分布に従うとする．

$$z_{dn}\sim\text{Multi}(\theta_d)$$
$$w_{dn}\sim\text{Multi}(\phi_{z_{dn}})$$

閲覧 Web サイトに関しても，番組視聴と同様に考える．トピック $k(=1,\ldots,K)$ におけるサイト $s(=1,\ldots,S)$ の出現確率を ψ_{ks} とする．$\psi_k=(\psi_{k1},\ldots,\psi_{kS})'$ とし，ψ_k の事前分布をパラメータ γ のディリクレ分布とする．γ は共通の要素 $\gamma_0(>0)$ からなるサイズ S のベクトルとする．

$$\psi_k\sim\text{Dirichlet}(\gamma)$$
$$\gamma_0\sim\text{LogNormal}(\mu_\gamma,\sigma_\gamma^2)$$

このように，ディリクレ分布のパラメータは対称である．また γ_0 の事前分布をパラメータ $\mu_\gamma,\sigma_\gamma^2$ に従う対数正規分布とする．γ_0 はデータから推定する．

消費者 d のサイト s の期間中の総閲覧回数を M_{ds} とする．すると $M_d=\sum_{s=1}^{S}M_{ds}$ は消費者 d の期間中の全サイトの総閲覧回数となる．

消費者 d の $m(=1,\ldots,M_d)$ 番目の閲覧 Web サイトにおけるトピックを y_{dm} とする．y_{dm} は離散値をとる潜在変数であり，パラメータ θ_d の多項分布に従うとする．また消費者 d の m 番目の閲覧 Web サイトを x_{dm} とする．x_{dm} はパラメータ $\psi_{y_{dm}}$ の多項分布に従うとする．

$$y_{dm}\sim\text{Multi}(\theta_d)$$
$$x_{dm}\sim\text{Multi}(\psi_{y_{dm}})$$

佐藤 [15] によれば，θ_d の事前分布は，パラメータ α について各 α_k が異なる，非対称 Dirichlet 分布に設定にしたほうが望ましい性質が多々あることが知られている．また，ϕ_k の事前分布と ψ_k の事前分布は，パラメータ β と γ のそれぞれの各要素が β_0 と γ_0 のように同じ値をとる，対称 Dirichlet 分布でもそれほど大差がないことが知られている．そこで本研究では，α は非対称で各 α_k は異なると想定し，β と γ に関しては対称で各要素は β_0 と γ_0 のように同じ値ととることとした．

データが得られたときの消費者 d の尤度 L_d と全体の尤度 L は以下のようになる．

$$L_d=\prod_{n=1}^{N_d}\left(\sum_{k=1}^{K}\theta_{dk}\phi_{kw_{dn}}\right)\cdot\prod_{m=1}^{M_d}\left(\sum_{k=1}^{K}\theta_{dk}\psi_{kx_{dm}}\right)$$
$$L=\prod_{d=1}^{D}L_d$$

このように，消費者 d の尤度は，番組視聴の各機会と Web サイト閲覧の各機会において，番組視聴と Web サイト閲覧行動に共通するパラメータ $\theta_d=(\theta_{d1},\ldots,\theta_{dK})'$ を潜在クラス確率とする尤度を計算し，これを番組視聴と Web サイト閲覧の全機会について掛け合わせたものである．そのため，視聴番組と Web サイト閲覧に関して，k が同じであれば，(ϕ_k,ψ_k) は同じトピックに属するものとして解釈することが可能となる．

モデルの推定はベイズ法により行う．推定では崩壊型ギブスサンプリングとメトロポリス・ヘイスティングス・アルゴリズムを用いた MCMC (Markov Chaine Monte Carlo) 法を用いる．

崩壊型ギブスサンプリングではまず z を，続いて y を以下の事後分布に従ってサンプリングする (岩田 [16])．

$$p(z_{dn}=k|W,X,Z^{\backslash dn},Y,\alpha,\beta,\gamma)\propto(N_{dk\backslash dn}+M_{dk}+\alpha_{dk})\frac{N_{k_{w_dn}\backslash dn}+\beta_0}{N_{k\backslash dn}+\beta_0V}$$

$$p(y_{dm}=k|W,X,Z,Y^{\backslash dm},\alpha,\beta,\gamma)\propto(N_{dk}+M_{dk\backslash dm}+\alpha_{dk})\frac{M_{k_{x_dm}\backslash dm}+\gamma_0}{M_{k\backslash dm}+\gamma_0S}$$

ただし，N_{dk} はギブスサンプリング中の消費者 d でのトピック k への割り当て回数，N_{kv} はギブスサンプ

リング中の番組 v でのトピック k への割り当て回数，$N_k = \sum_{v=1}^{V} N_{kv}$ はギブスサンプリング中の番組視聴でのトピック k への割り当て回数，M_{ks} はギブスサンプリング中のサイト s へのトピック k への割り当て回数，$M_k = \sum_{s=1}^{S} M_{ks}$ はギブスサンプリング中のサイト閲覧でのトピック k への割り当て回数，$A\backslash B$ は A のうち B 以外の要素，$\backslash C$ は C を除く全ての要素，である．

α, β, γ については，メトロポリス・ヘイスティングス・アルゴリズムでサンプリングを行う．

$$p(\alpha_k|\alpha_{\backslash k}, W, X, Z, Y, \beta, \gamma) \propto p(\alpha_k)p(Z, Y|\alpha)$$

$$p(\beta_0|W, X, Z, Y, \alpha, \gamma) \propto p(\beta_0)p(W|Z, \beta)$$

$$p(\gamma_0|W, X, Z, Y, \alpha, \beta) \propto p(\gamma_0)p(X|Y, \gamma)$$

3. 利用データとモデルの推定

3.1 利用データの概要

本研究では実証分析として，平成 30 年度データ解析コンペティションで貸与された株式会社ビデオリサーチ『VR CUBIC』のメディア接触データを利用した．データ期間は 2017 年 4 月 3 日（月）～2018 年 4 月 1 日（日）である．

テレビ番組の分析対象としてドラマを選択した．木村ら [17] による 2015 年の調査では，ふだんよく見る番組は上位から「ニュース・ニュースショー・報道番組（76%）」，「天気予報（53%）」，「ドラマ（50%）」であり，ドラマはふだんからよく見られる番組であり，放送局にとって重要な番組である．さらに，ドラマはジャンルが多岐にわたり消費者の好みや価値が視聴行動に反映されることを期待できる．以上の理由から，メディア消費への「モチベーション」を探る今回の研究の対象として適切であるといえる．

分析対象者はデータ期間中に「ドラマ番組の視聴が 10 回以上」かつ「Web ページ閲覧が 10 回以上 2,000 回以下」の 795 名とした．ドラマは複数回の放送がなされており，また同じ回のドラマが複数の時間帯で放送されることもあるが，データセットに割り振られた番組コードにより番組を区別した．消費者個人別に 1 日あたり 10 分以上の視聴があれば 1 とカウントした．またリアルタイムとタイムシフト視聴は同じ番組視聴として区別をしなかった．分析対象ドラマは 505 番組であった．Web サイトの閲覧に関してはサブドメイン単位で 1 日のうち 10 秒以上閲覧があれば 1 とカウントした．さらに分析対象 Web サイトは閲覧者数が分析対象者中 50 人以上のものに限った．この結果，分析対象 Web ページは 441 サブドメインとなった．

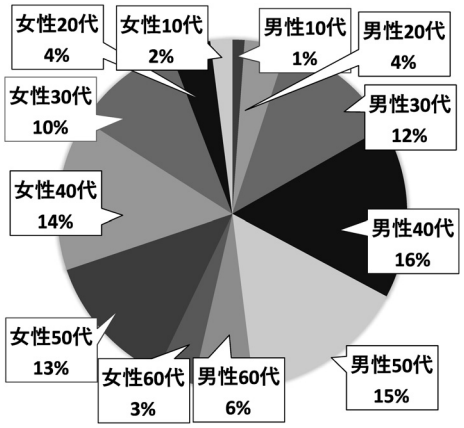

図 2　分析対象者の性別年齢別の分布

図 2 は分析対象者の性別年齢別の分布である．分析対象者は男性が 53.6%であり，男性のほうが女性よりもやや多い．年齢では男女ともに 40 代が最も多い．

図 3 は TV ドラマの総視聴回数とその順位，および，Web サイトの総閲覧回数とその順位である．順位と回数のスケールはそれぞれ常用対数である．もし順位と回数の関係が冪乗則に従う場合には，図 3 において両者の関係は直線になることが期待される．TV ドラマにおいては視聴回数が上位の番組ほど直線から外れていることから，視聴回数の多い上位番組に消費者の視聴が分散していることがわかる．一方 Web サイトに関しては，上位 2 つのサイトは閲覧回数が拮抗しているが，順位が 10 位以降のサイトでは直線に近く，順位と総閲覧回数の関係は冪乗則に近い傾向にあることがわかる．このようにドラマと Web サイトでは集計的な消費行動においても，構造的な差があることがわかる．

3.2 モデルの推定とトピック数の決定

モデルの推定には，崩壊型ギブスサンプリングとメトロポリス・ヘイスティングス・アルゴリズムを用いた MCMC 法によりベイズ推定を行った．MCMC 法では 20,000 回のサンプリングを行い，後半 10,000 サンプリングのうちの 10 サンプリングに 1 回をモデルパラメータの事後分布として利用した．

推定のためには，アプリオリにトピック数を与えることが必要である．ジョイント LDA モデルのトピック数を決定する前に，まずは番組視聴のみを考慮した番組 LDA モデルについて，トピック数を 2 から 10 の間で

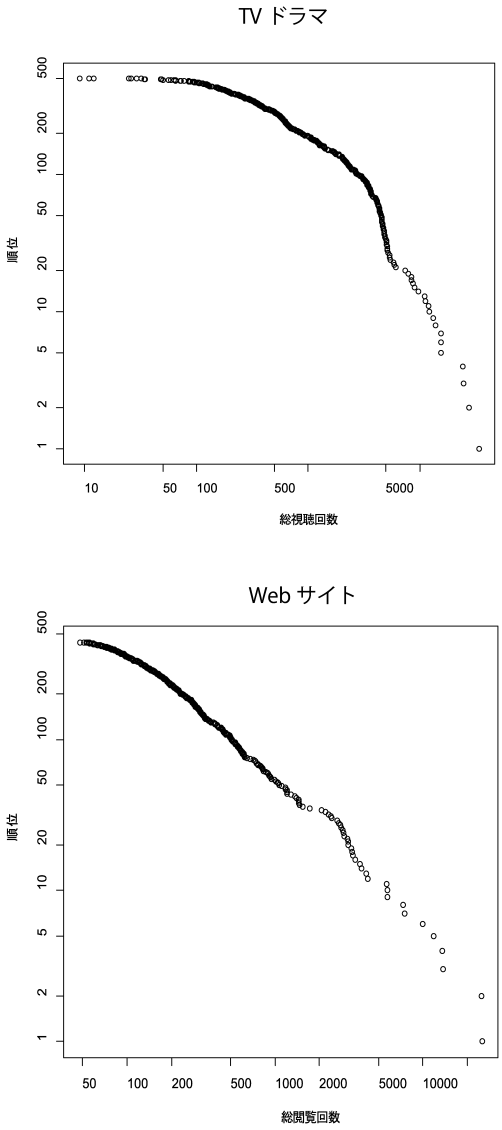

図 3　総視聴（閲覧）回数と番組（サイト）順位の関係

間隔 1 で変化させて対数周辺尤度を比較した（図 4 の上）．対数周辺尤度が最も高くなるのはトピック数が 5 の場合であった．一方，Web サイト LDA モデルについて，トピック数を 2 から 10 の間で間隔 1，その後は 15, 20, 30 と変化させて対数周辺尤度を比較した（図 4 の中）ところ対数周辺尤度はトピック数が 9 で一度減少し，その後は上昇した．最後にジョイント LDA モデルについて，トピック数を 2 から 15 の間で間隔 1 で変化させ，その後 20 まで増やしたとき，対数周辺尤度はトピック数が 2 の時に最大となった．番組 LDA モデルでのトピック数は 5 であったため，結果の解釈の有益性の観点から，番組 LDA モデルのトピック数よりも多いトピック数に限ってジョイント LDA モデルのトピック数を検討し，トピック数が 10 で対数周辺尤度が最大となったため，ジョイント LDA モデルのトピック数は 10 に決定した．

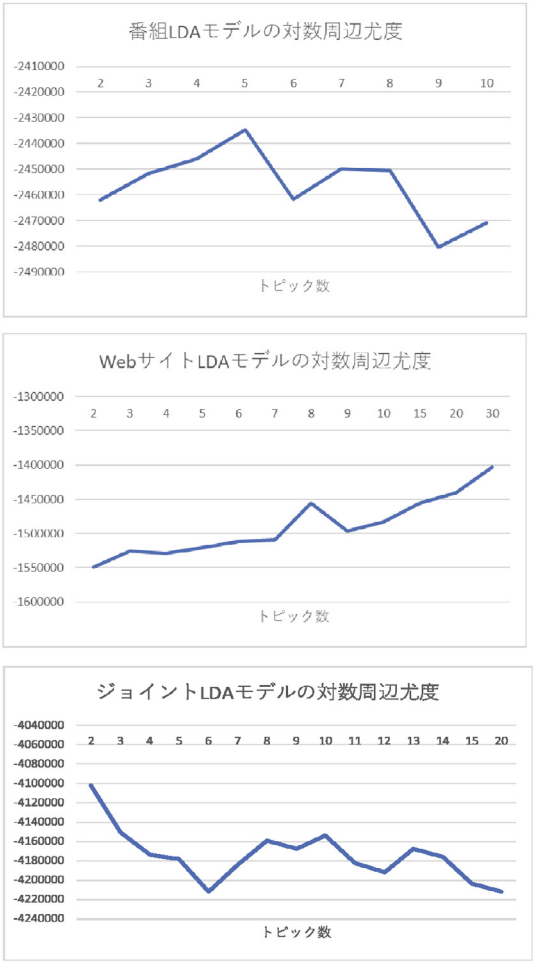

図 4　対数周辺尤度の比較

4.　ジョイント LDA モデルによる分析結果

4.1　メディア消費の統合的分析

先の 3.2 節でジョイント LDA モデルではトピック数を 10 に決定した．表 1 は各トピックにおける，トピックの比率（シェア）と性別年齢別の構成比である．性別年齢の変数はモデル構造に含まれないため，トピック毎に事後的に集計を行った．トピック 2 とトピック 6 で全体の 44.7% を占める．最も男性の比率が高いトピックはトピック 1 であり男性比率が 71.1% である．一方，最も女性の比率が高いトピックはトピック 9 であり，女性比率が 64.3% を占める．

各トピックの特徴は，トピック k での視聴機会毎の番組の視聴確率 ϕ_k と閲覧機会毎のサイトの閲覧確率

ψ_k をもとに解釈することができる．各トピックの上位 20 位までの ϕ_k と ψ_k をもとにトピックの特徴をまとめた結果と，ジョイント LDA モデルの結果をもとに事後的に消費者を集計して得られた性別年齢の特徴は以下のとおりである．なお最後の括弧内の数値はトピックの比率である．

トピック 1：時代劇視聴，古くからある Web サイトのユーザー．40 代以上男性が多い．（2.3%）

トピック 2：プライムタイム視聴，Google の検索サービスとメールを利用．30 代以下が多い．（22.4%）

トピック 3：刑事・サスペンスドラマ視聴，Web でポイント収集．50 代以上男性が多い．（8.1%）

トピック 4：帯ドラマ視聴，Web で動画鑑賞と交流．40 代以下が多い．（7.5%）

トピック 5：朝ドラ・韓流ドラマ視聴，Web はオークションとファイナンス利用．40 代以上が多い．（6.9%）

トピック 6：刑事・サスペンスドラマ視聴，Yahoo のニュースと検索サービスを利用．30 代以上が多い．（22.4%）

トピック 7：帯ドラマ，朝ドラマ視聴，Web でショッピング．50 代以上が多い．（5.6%）

トピック 8：週末 TV 視聴，Yahoo のメールとショッピングを利用．20 代以上が多い．（8.7%）

トピック 9：再放送視聴，複数の Web 検索サービスを利用．女性が多い．（8.2%）

トピック 10：深夜ドラマ視聴，楽天ユーザー．30 代から 50 代が多い．（7.9%）

なお，TV 番組の視聴行動データのみを利用した番組 LDA モデルでのトピックの特徴は以下のようになった（トピック数は対数周辺尤度から 5 に決定）．

トピック 1：再放送視聴．40 代以上女性が多い．（14.7%）

トピック 2：刑事ドラマ・プライムタイム視聴．全年代に分布．（56.8%）

トピック 3：連続ドラマ・帯ドラマ視聴．50 代以上女性が多い．（6.9%）

トピック 4：朝ドラ・韓流視聴．30 代以上が多い．（10.1%）

トピック 5：早朝・深夜ドラマ視聴．30 代以上男性が多い．（11.6%）

このように TV 番組のみを利用して分析を行うと，トピックは放送時間帯によって分かれることがわかる．これに対し，ジョイント LDA モデルで TV 番組視聴に Web サイト閲覧を加えると，時間帯以外の好みや興味の要因が加わる．このため，番組 LDA モデルでのトピックが分割・再構成されることが期待できる．

実際，番組 LDA モデルではトピック 2 の刑事ドラマ・プライムタイムは，ジョイント LDA ではトピック 2・3・6・8 に分割されて，視聴番組が細分化され，またそれぞれ閲覧 Web サイトが異なっている．一方，トピックの比率は小さくなっているが，番組 LDA モデルではトピック 1 の「再放送視聴」はジョイント LDA ではトピック 9 の「再放送視聴，複数の Web 検索サービスを利用」に対応している．このように，番組 LDA モデルからジョイント LDA モデルへ類似するトピックを対応させた場合，番組 LDA モデルのトピックの分割のされ方がトピックによって異なることからも，ジョイント LDA モデルでは視聴番組と閲覧 Web サイトの関連性が考慮されていることがわかる．もし番組

表 1　各トピックでの性別年齢別構成比率

		トピック1	トピック2	トピック3	トピック4	トピック5	トピック6	トピック7	トピック8	トピック9	トピック10
トピックの比率		2.3%	22.4%	8.1%	7.5%	6.9%	22.4%	5.6%	8.7%	8.2%	7.9%
性別年齢別構成比	男性10代	0.2%	2.4%	0.3%	2.1%	1.1%	0.6%	0.1%	0.7%	0.7%	0.7%
	男性20代	2.0%	4.6%	1.5%	9.7%	1.8%	3.5%	0.9%	3.8%	3.2%	5.4%
	男性30代	9.5%	13.9%	7.7%	13.5%	11.4%	12.6%	6.2%	14.1%	7.4%	12.0%
	男性40代	28.6%	12.7%	17.4%	11.7%	13.7%	19.1%	12.3%	20.2%	11.4%	21.5%
	男性50代	22.4%	13.3%	20.0%	14.7%	14.3%	14.6%	21.2%	17.2%	7.7%	17.9%
	男性60代	8.3%	4.7%	12.3%	5.2%	7.5%	2.7%	8.0%	6.1%	5.2%	4.9%
	女性10代	0.2%	3.5%	1.7%	2.0%	1.0%	1.3%	0.9%	0.9%	3.4%	2.0%
	女性20代	0.7%	5.2%	2.7%	3.0%	4.2%	2.7%	2.2%	3.6%	7.0%	3.4%
	女性30代	8.9%	10.9%	7.4%	7.1%	12.4%	9.8%	10.9%	8.4%	14.2%	10.6%
	女性40代	7.1%	16.0%	16.4%	11.4%	13.9%	12.9%	14.4%	12.0%	20.0%	12.1%
	女性50代	11.5%	10.2%	10.5%	16.1%	15.9%	15.6%	14.7%	10.0%	14.4%	8.2%
	女性60代	0.5%	2.6%	2.1%	3.6%	3.0%	4.6%	8.3%	3.0%	5.3%	1.2%

LDA モデルと Web サイト LDA モデルを独立に推定して結果を掛け合わせた場合には，モデル間でトピックは独立しているため，ジョイント LDA モデルのように視聴番組と閲覧 Web サイトの関連性を見出すことが難しくなる．ジョイント LDA モデルでは，メディア間の関連性はモデル構造として考慮されているため，視聴番組と閲覧 Web サイトの関連性を把握することが可能となるのである．

さらに，ジョイント LDA モデルでは視聴番組と閲覧サイトの特徴を同時に解釈することで，消費者のメディア消費へのモチベーションを創発することができる．例えば，トピック 3 とトピック 6 では刑事・サスペンスドラマ視聴であるが，Web サイト閲覧においてはトピック 3 ではポイント収集のような手間をかけて金銭的報酬を得ることを動機としており，トピック 6 では Yahoo のニュースや検索サービスの利用のような情報収集を動機としている．Austin [18] では映画館へ出かけるモチベーションについて 12 種類のモチベーションを特定しているが，これを本研究での結果にあてはめて考えると，ジョイント LDA のトピック 3 は「時間つぶし」であり，トピック 6 は「会話の話題集め」である．このように，TV 番組の視聴だけから区別することができないメディア消費へのモチベーションを，Web サイトの閲覧を加えることで，その特徴をうまく抽出することができた．

4.2　TV 番組と Web サイトの潜在的利用率の評価

次にジョイント LDA モデルを利用して TV 番組や Web サイトの潜在的利用率の評価を行う．これは番組や Web サイトの潜在的な共起関係から「視聴可能性の高い番組」と「閲覧可能性の高い Web サイト」を抽出するものである．計算方法については里村 [6] と同じ方法を用いた．

消費者 d の番組 v の視聴確率の予測値 $\Pr(w_d = v)$ とサイト s の閲覧確率の予測値 $\Pr(x_d = s)$ は以下の式から求める．

$$\Pr(w_d = v) = \sum_{k=1}^{K} p(v|k)p(k|d) = \sum_{k=1}^{K} \phi_{kv}\theta_{dk}$$

$$\Pr(x_d = s) = \sum_{k=1}^{K} p(s|k)p(k|d) = \sum_{k=1}^{K} \psi_{ks}\theta_{dk}$$

すべての消費者について予測値を求めた後，番組 v の視聴者の中から視聴確率予測値の 50%点を求め，番組 v の未視聴者の中で，この値より大きい番組 v の視聴確率予測値を持つ消費者を番組 v の潜在的視聴可能性の高い消費者とした．同様に，サイト s の閲覧者の中から閲覧確率予測値の 50%点を求め，サイト s の未閲覧者の中で，この値より大きいサイト s の閲覧確率予測値を持つ消費者をサイト s の潜在的閲覧可能性の高い消費者とした．

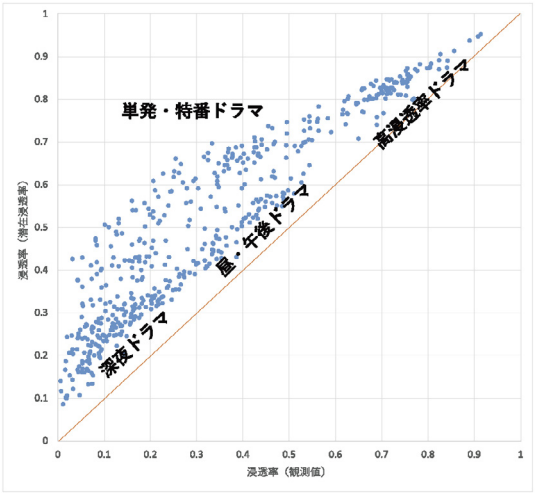

図 5　TV 番組の潜在的視聴の可能性

図 5 は各番組の視聴可能性の高い消費者の比率（潜在浸透率）を計算したものであり，図 6 は各 Web サイトの閲覧可能性の高い消費者の比率（潜在浸透率）を計算したものである．各図ともに，横軸は各番組を一度でも視聴あるいは各 Web サイトを一度でも閲覧したこのとある消費者の比率（浸透率の観測値），縦軸は潜在浸透率である．なお潜在浸透率の計算では番組の既存視聴者と Web サイトの既存閲覧者も含めているため，各点は 45 度対角線よりも上に付置される．

図 5 の番組の潜在浸透率を見ると，多くの番組は 45 度対角線上に近く，これ以上の浸透可能性は高くないことがわかる．特に，観測値での浸透率が高い「高浸透率ドラマ」はそもそも浸透率が高いために，これ以上の視聴者を増やすことが難しいことがわかる．一方，浸透率が中程度の「昼・午後ドラマ」や，浸透率が低い「深夜ドラマ」は，時間帯の制約もあるため，現在の視聴者を超えて他の視聴者へ浸透させることが難しいと解釈できる．一方，図 5 の左上には潜在浸透率が高い番組として「単発・特番ドラマ」がある．これらの番組は，放送回数が他の番組と比べて少ないために観測値での浸透率が低くなっていると考えられる．そこで単発・特番ドラマについては，放送前の番組の宣伝などにより認知を促進することが，観測値の浸透率を伸ばすための施策として考えられる．

次に図 6 の Web サイトの潜在浸透率を見ると，現在

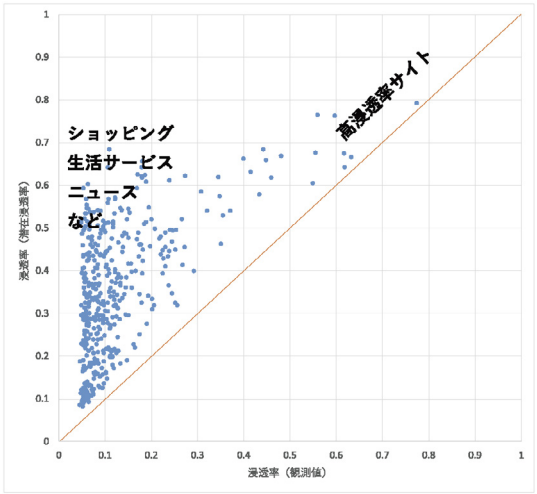

図 6　Web サイトの潜在的閲覧の可能性

の浸透率が低くても，浸透の可能性が高いと評価されたサイトがある．それらのサイトはショッピング，生活サービス，ニュースなどである．これらのサイトは現在の利用がなくても消費者に利用してもらえる可能性が高いサイトであるといえる．一方，浸透率が既に高いサイトは潜在的浸透率がそれより高くなることが難しいことがわかる．

4.1 節ではテレビ番組は放送時間帯によりトピックが決まっていることがわかったが，潜在浸透率に関する分析でも，テレビ番組に関しては，既に視聴する消費者が放送時間帯によって固定化しており，そのような番組はこれ以上の浸透を行うことは難しいことがわかった．一方，Web サイトに関しては，サイトへの閲覧者の固定化の程度は弱く，多くのサイトが新しい消費者に閲覧してもらえる可能性があるといえよう．

5.　おわりに

本研究ではジョイント LDA モデルを利用して TV 視聴データと Web サイト閲覧データを結びつける手法の提案を行った．提案手法は TV と Web サイトの利用行動を同時に分析することで統合的な消費者インサイトを獲得することを目指すものである．

実証分析の結果，TV 番組の視聴行動は放送時間帯の制約を大きく受けていることがわかった．ジョイント LDA モデルでは，TV 番組の視聴行動と Web サイトでの閲覧行動を，メディア間の関連性も考慮して分析することで，特徴のあるトピックを抽出することができた．また，浸透率を伸ばせる TV 番組は単発・特番ドラマであり，浸透率を伸ばせる Web サイトはショッピング・生活サービス・ニュースであることが示された．

最後に本研究の課題と今後の研究の可能性について述べたい．番組 LDA モデルでの分析では，トピックは放送時間帯によって決まっていた．このような結果が得られた理由として，視聴者は各自の視聴可能な時間帯の中で番組選択を行い，放送局は各時間帯の視聴者層を予想しながら番組編成を行っていることが挙げられる．TV 番組の視聴データの分析において，このような内生性の問題を考慮することは，今後の研究の課題である．また，Web サイト閲覧については，閲覧情報をさらに活用することが考えられる．例えば閲覧時間帯や閲覧継続時間の情報を利用することで，さらなる示唆を得ることが期待される．

謝辞　本研究の分析では「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」「株式会社ビデオリサーチ　VR CUBIC」から提供されたデータを使用しました．関係者各位に感謝の意を表します．

参考文献

[1] 総務省情報通信政策研究所，「平成 29 年情報通信メディアの利用時間と情報行動に関する調査報告書」，https://www.soumu.go.jp/main_content/000564530.pdf（2019 年 5 月 6 日閲覧）
[2] J. Büschken and G. M. Allenby, “Sentence-based text analysis for customer reviews,” *Marketing Science*, **35**(6), pp. 953–975, 2016.
[3] B. J. D. Jacobs, B. Donkers and D. Fok, “Model-based purchase predictions for large assortments,” *Marketing Science*, **35**(3), pp. 389–404, 2016.
[4] M. Trusov, L. Ma and Z. Jamal, “Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting,” *Marketing Science*, **35**(3), pp. 405–426, 2016.
[5] A. Ansari, Y. Li and J.Z. Zhang, “Probabilistic topic model for hybrid recommender systems: A stochastic variational bayesian approach,” *Marketing Science*, **37**(6), pp. 987–1008, 2018.
[6] 里村卓也，“トピックモデルによる顧客データの統合的分析，”オペレーションズ・リサーチ：経営の科学，**63**(2), pp. 67–74, 2018.
[7] D. M. Blei and M. I. Jordan, “Modeling annotated data,” In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 127–134, 2003.
[8] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith and A. McCallum, “Polylingual topic models,” In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, **2**, pp 880–889, 2009.
[9] T. Iwata, S. Watanabe and H. Sawada, “Fashion coordinates recommender system using photographs

from fashion magazines," In *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI*, pp. 2262–2267, 2011.
[10] S. Pyo, E. Kim and M. Kim, "LDA-based unified topic modeling for similar TV user grouping and TV program recommendation," *IEEE Transaction on Cybernetics*, **45**(8), pp. 1476–1490, 2015.
[11] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, **3**, pp. 993–1022, 2003.
[12] P. Arabie and L.J. Hubert, "Cluster analysis in marketing research," *Advanced Methods in Marketing Research*, R. P. Bagozzi (ed.), Blackwell, pp. 160–189, 1994.
[13] 岡太彬訓，守口剛，『マーケティングのデータ分析―分析手法と適用事例―』，朝倉書店，2010.
[14] 黒木学，山下遥，"改良型 k-planes クラスター分析法と解析結果の視覚化について，"日本経営工学会論文誌，**68**(1), pp. 1–12, 2017.
[15] 佐藤一誠，『トピックモデルによる統計的潜在意味解析』，コロナ社，2015.
[16] 岩田具治，『トピックモデル』，講談社，2015.
[17] 木村義子，関根智江，行木麻衣，"テレビ視聴とメディア利用の現在―『日本人とテレビ・2015』調査から―，"放送研究と調査，**65**(8), pp. 18–47, 2015.
[18] B. A. Austin, "Motivations for movie attendance," *Communication Quarterly*, **34** (2), pp. 115–126, 1986.