潜在表現モデルに基づくテレビ番組の魅力度要因分析に関する研究

1X15C083-2　西村 祐樹
指導教員　後藤 正幸

1　研究背景と目的

近年，テレビ業界において，情報収集や娯楽のためにテレビ番組を選択する人が相対的に減少する “テレビ離れ”が問題視されている．テレビ離れによる視聴率の低下は各放送局の収入の減少に繋がるため，各放送局では，視聴者にとって魅力的な番組を制作し，テレビ視聴を促進することが極めて重要となっている．

これに伴い，各放送局は魅力的な番組制作を考える必要がある．魅力的なテレビ番組を制作するためには，出演者，番組内容，裏番組（対象番組と同日同時刻に放送されている番組）といった要因が視聴率に与えるインパクトを評価できるモデルがあると大変有用である．このモデルを有効活用し，様々な魅力度要因の影響を分析することで，新たな番組制作に役立てることが期待できる．

そこで本研究では，インターネット上で取得可能な各番組の「出演者情報」と「番組情報」のテキストデータ，並びに同時間帯における全番組の「視聴率」データを用い，これらの関係性を表現する分析モデルを提案する．その際，これらの要因間の非線形な関係性をモデルに取り込みつつ，過学習を防ぐため，高次元の「出演者情報」は積層自己符号化器 (以下，SAE) [1] を適用した次元圧縮を，テキストデータである「番組情報」は単語ベクトル化の後に潜在的ディリクレ配分法 (Latent Dirichlet Allocation，以下，LDA) [2] を適用して低次元トピックベクトル化を行う．そして，これらを入力として順伝播型ニューラルネットワーク (以下，FNN) [3] モデルを学習させる．

本研究では，実データを用いてモデルを構築し，提案モデルの予測性能について妥当性を検証する．また，得られたFNN モデルを用い，様々な「出演者情報」,「番組情報」を変化させた場合のインパクトを評価することで，最適な出演者や番組内容の決定といった，魅力度向上のための施策考案が可能であることを示す．

2　提案モデル

本研究では，非線形な関係性を表現できる FNN をベースとして，テレビ番組の魅力度と番組構成要素・裏番組などの影響要因間の複雑な関係性をモデル化する．そして，得られたモデルを用いて，魅力度を向上させる施策を検討する．

(1) 対象データ

提案モデルでは，各番組の出演者情報，番組情報，視聴率情報を扱う．まず，各番組の出演者情報は，出演者の出演の有無を 1,0 のダミー変数で表しているベクトルである．また，番組情報は，インターネット上に公開されているテレビ番組内容を記述したテキストデータである．これらの番組情報は，形態素解析によって得られた単語を用いて，出現頻度でベクトル化した単語頻度ベクトルに変換して用いる．さらに，各番組の視聴率情報は，各番組の 1 分ごとの視聴率からなるデータである．

(2) 目的変数

テレビ番組の魅力度を表す指標として，一般的な視聴率が広く受け入れられやすいと考えられる．しかし，各番組の視聴率は，裏番組に多大な影響を受けるため，各番組の視聴率をその番組の特徴量のみから予測するモデルの構築は困難である．そこで，同日同時刻に放送されている全番組の特徴量からそれらの番組の視聴率を同時に予測することで，裏番組を考慮したモデルの構築を行う．さらに本研究では，曜日や時間帯などの，全番組に共通する要因の影響を除去するため，同日同時刻の各番組の視聴率をそれらの番組の視聴率の和で除した “視聴割合” を目的変数として用いる．これにより，各番組の固有な構成要素が与えるテレビ番組の魅力度への影響に着目することができる．すなわち，各番組の視聴割合を魅力度と定義し，同日同時刻 (1 分単位) に放送されている番組の視聴割合を連結させたベクトルを FNN の目的変数として用いる．

(3) 説明変数

本研究では，出演者情報と番組情報が各番組固有のテレビ番組の魅力度に影響する要因であると考える．しかし，出演者情報や単語頻度ベクトルは高次元でスパースなデータである．故に，そのまま FNN の入力とすると，過学習によりモデル化が上手く行えない可能性がある．この点を考慮し，本研究ではこれらのデータに対し，予め次元圧縮を行う．

出演者情報に対しては，非線形次元圧縮を行う．代表的な非線形次元圧縮手法としてカーネル主成分分析があるが，カーネル主成分分析ではグラム行列の固有値計算を行うため，計算コストがかかってしまう．そこで，本研究では出演者情報に対し，非線形データの効率的な圧縮が可能な SAE を適用し，圧縮表現（以下，出演者ベクトル）を獲得する．

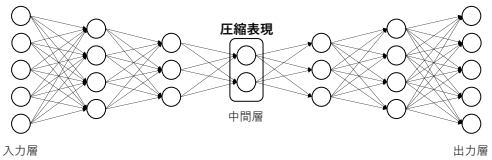

図 1: SAE による圧縮

一方，単語頻度ベクトルに対しても SAE を適用することが考えられる．しかし，テキストデータに対しては，単語の生起が確率的な揺らぎを持っていると仮定する統計モデルの方が当てはまりが良いことが数々の研究で明らかになっている．そこで，本研究では単語頻度ベクトルに対し，頻繁に用いられる統計モデルである LDA を適用し，圧縮表現（以下，トピックベクトル）を獲得する．

(4) FNN モデルの構築

前述の説明変数と目的変数を用いて FNN の学習を行う．ただし，出演者ベクトルとトピックベクトルは，FNN の学習でどちらかの影響要因だけが重視されないよう，同じ次元数に圧縮されているとする．そして，各番組に関する出演者ベクトルとトピックベクトルを連結したベクトルを各番組固有の特徴を表す番組表現ベクトルと定義する．さらに，目的変数の順に同日同時刻に放送された各番組の番組表現ベクトルを連結させたベクトルを FNN の説明変数として用いる．これにより，裏番組を考慮した魅力度要因分析を実現する．

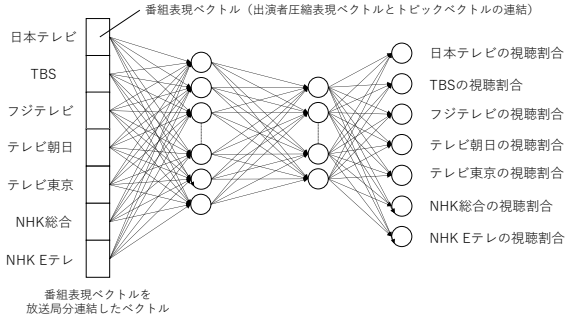

図 2: FNN モデルの構成

(5) 魅力度要因分析

学習済み FNN モデルを用いることで魅力度要因分析を行う．具体的には，番組表現ベクトルや出演者ベクトルなどの説明変数を変化させ，視聴割合への影響を観察する．

[提案モデルの構築手順]

STEP1： SAE を用いた出演者情報の次元圧縮

各テレビ番組の出演者情報を入力とした SAE の学習を行う．学習済みの SAE に対象の出演者情報を入力することで，圧縮された出演者ベクトルを獲得する．

STEP2： LDA を用いた番組情報の次元圧縮

各テレビ番組の番組情報から単語頻度ベクトルを作成し，それらを LDA の入力とすることで，テレビ番組内容を考慮したトピックベクトルを獲得する．

STEP3： FNN の学習

STEP1-2 で獲得した圧縮表現を入力として，FNN を学習する．

3 実データ分析

3.1 分析条件

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関東 1 都 6 県のインターネット利用者約 5,000 人を対象としたテレビ視聴データ (株式会社ビデオリサーチの VR CUBIC データ) 及び，Web 上より入手した出演者データと番組情報データを提案モデルに適用する．対象データの期間は 2017 年 4 月 3 日から 2018 年 4 月 1 日，データ件数は 524,160 件 (対象データ期間の分数と一致)，出演者数は 9,562 名，分析対象のテレビ番組は主要放送局 7 社 (図 2 参照) が放送している番組とする．また，事前実験により，SAE は 7 層，ユニット数は入力層から順に 9,562，3,000，500，30，500，3,000，9,562 とし，LDA のトピック数は 30 とする．FNN のユニット数は入力層から順に 420，300，100，7 とする．この時，SAE より獲得する 30 次元の出演者ベクトル及び LDA より獲得する 30 次元のトピックベクトルを連結させた 60 次元の番組表現ベクトルを 7 社分連結させた 420 次元のベクトルが FNN の入力ベクトルとなる．また，出演者情報や番組情報が存在しない番組に対しては，その番組表現ベクトルを **0** ベクトルとする．

3.2 分析結果と考察

(1) 提案モデルの正当性

まず，学習データ：テストデータをランダムに 8 : 2 に分割して学習を行った．SAE，LDA，FNN それぞれの性能を示す評価指標とその値を表 1 に示す．

表 1: 各モデルの学習後の性能評価値

モデル	学習データ	テストデータ	評価指標
SAE	0.0099	0.0042	交差エントロピー
LDA	0.4713	1.0155	perplexity ($\times 10^3$)
FNN	0.2045	0.2051	交差エントロピー

表 1 の結果より，提案モデルにおいてテストデータに対しても頑健な各番組間の視聴割合を予測するモデルが構築されていることがわかる．

(2)**LDA で推定された潜在的なトピックの解釈**

LDA で推定した番組トピックへの解釈を表 2 に示す．

表 2: LDA により推定したトピックの解釈

番号	トピック	番号	トピック
1	散歩	16	世界観光
2	時代劇	17	バラエティー（お笑い）
3	料理	18	政治
4	バラエティー（情報）	19	出演者
5	教育	20	アイドル・男性タレント
6	芸術	21	レジャー
7	小学生向け	22	母親向け
8	犯罪・サスペンス	23	製作者
9	ニュース（天気）	24	生物
10	ニュース（総合）	25	歌・大学
11	バラエティー（クイズ）	26	国内観光
12	アイドル・女性タレント	27	ニュース（社会）
13	自然	28	バラエティー（ドキュメンタリー）
14	挑戦・オリンピック	29	乳幼児向け
15	医療	30	道徳・24 時間テレビ

表 2 のように，各テレビ番組の潜在的なトピックに対しそれぞれ異なった解釈を与えることができる．また，一般的な番組のジャンルよりも詳細なカテゴリを推定できていることがわかる．

(3) 提案モデルを用いた分析

学習済みモデルを用いて，出演者の追加とテレビ番組を変化させた時の視聴割合の変動を分析する．ここでは，事例としてフジテレビで 2017 年 6 月 8 日の 21 時 15 分に放送されていた「とんねるずのみなさんのおかげでした」(以下，対象番組) を取り上げる．対象番組における出演者を表 3 に，上位 5 つの所属トピックを表 4 に示す．

表 3: 出演者一覧

出演者
とんねるず
カミナリ
マッコイ斉藤
小倉優子
小木博明

表 4: 上位 5 所属トピック

順位	トピック
1	挑戦・オリンピック
2	全国スポット紹介
3	バラエティー（ドキュメント）
4	バラエティー（お笑い）
5	道徳・24 時間テレビ

また，学習後の FNN を用いて，対象番組に出演者を 1 人追加した際のインパクトを評価したところ，表 5 に示す出演者を追加した場合に対象番組の予測視聴割合が上昇するという推定結果が得られた．

表 5: 対象番組の予測視聴割合を増加させる追加出演者上位 5 名

出演者	予測視聴割合	増加視聴率
国分太一	18.01%	0.06%
渡部建	17.94%	0.05%
阿部渉	17.93%	0.05%
船越英一郎	17.92%	0.04%
森富美	17.88%	0.04%
追加なし	17.69%	-

また，対象番組を同放送局の他番組と置き換えたところ，表 6 に示す番組で予測視聴割合が上昇する結果となった．

表 6: 番組変更の結果，予測視聴割合を増加させる番組上位 5 件

番組	予測視聴割合	増加視聴率
「村上信五とスポーツの神様たち」	28.50%	2.07%
「ROAD TO EDEN」	28.02%	1.97%
「めちゃ×２イケてるッ！記録より記憶に残る名作ベスト１０」	26.72%	1.73%
「ワイドナショー」	26.61%	1.70%
「文書改ざんキーマン佐川前長官を証人喚問みんなのニュース SP」	26.11%	1.61%

視聴割合を増加させる追加出演者（表 5）は，対象番組のようなバラエティー番組に出演しているタレントが多く含まれており，提案モデルは出演者と番組トピックの相性を考慮できていると考えられる．また，視聴割合を増加させる番組（表 6）には，バラエティー（クイズ），犯罪・サスペンス，自然，芸術，時代劇などのトピックが多く含まれている．これらのトピックには，裏番組に含まれているトピックと裏番組に含まれていないトピックが混在していた．そのため，提案モデルは，対象時間帯において視聴者に支持されているトピックや，新しい可能性を持つトピックの発見に活用できると考えられる．

4 まとめと今後の課題

本研究では，出演者情報を SAE で，番組内容のテキストデータを LDA で圧縮して低次元表現し，それらを用いて裏番組を考慮した視聴割合を予測する FNN モデルを提案した．また，提案モデルを用いて出演者の追加や番組を変化させた際の効果を分析した．提案モデルは裏番組を考慮しつつ，視聴者が魅力を感じるテレビ番組の内容や出演者について分析を行えるため，魅力的な番組制作への活用が期待できる．

今後の課題として，別条件での視聴割合の変動分析，流行要因の取り入れ，対象放送局数の追加，提案モデルのハイパーパラメータを適切に調整することなどが挙げられる．

参考文献

[1] G. Hinton, R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313. no. 5786, pp. 504 - 507, 2006.

[2] D. Blei, A. Ng, M. Jordan, “Latent Dirichlet Allocation,” *The Journal of Machine Learning Research*, Vol. 3, pp.993-1022, 2003.

[3] D. Rumelhart, G. Hinton, R. Williams, “Learning internal representations by error propagation,” *Parallel Distributed Processing*, Vol.1, chapter 8, pp.318-362, 1986.