CO-003

大規模テレビ視聴データを用いた行動学的属性に基づくセグメント抽出手法
A Segment Extraction Method Based on Behavioral Attributes Using Large Scale TV Viewing Data

田之上 伸吾† 大盛 善啓†
Shingo Tanoue Yoshihiro Ohmori

1. はじめに

2017年の日本の総広告費は6兆3,907億円であり、6年連続で緩やかに増加している[1]。媒体別に見ると、テレビ広告市場が1兆9,478億円、構成比30.5%と最も大きく、インターネットやモバイル通信の普及などのメディア環境の変化が著しい現代においても、テレビは依然として主要な広告媒体である。テレビ広告は主に、広告主が自社ブランドの認知率を向上するために出稿する広告と、放送局が自局の番組の視聴率を上げるために出稿する広告の 2 つに大別される。特に後者は番組宣伝広告（番宣）と呼ばれ、 視聴者に対して数ある商品・サービス・番組の中から好みの番組を視聴する機会を提供したり、放送局に対して自局番組の視聴率増加や広告収入増加に貢献したり、広告主に対してブランド認知増加や売上増加に貢献したりできる。よって、番宣を当該番組に興味がありそうな視聴者に効率よく接触させることは、たいへん重要な課題である。

ターゲットとなる視聴者集合（セグメント）を定める際、テレビ業界では人口統計学的属性（demographic attributes）を用いることが多い。例えば、表 1 に示すように、性別や年齢によって視聴者をグループ分けすることがよく行われる[2]。しかし、視聴者の趣味趣向やライフスタイルが多様化した現代において、人口統計学的属性で視聴者を適切に絞り込むことが困難なケースが増えてきている。例えば、子供向けに制作されたアニメ番組が、実は30代男性から高い支持を得ているケースなどが存在する。そのため、インターネット広告の分野で成功を収めている、検索履歴や閲覧履歴を用いた行動ターゲティング広告のような、行動学的属性（behavioral attributes）に基づくセグメント抽出をテレビに適用することが注目されている[3]。

これまでパネル調査によって得られた代表データを用いることが一般的であったが、データのボリュームや精度に課題があった。サンプル数と標本誤差の関係の例を表 2 に示す[2]。例えば、パネルを募集・保持するにはコストがかかるため、関東地区におけるパネル調査で得られるデータのサンプル数は数千程度である[4]。仮に 2,000 サンプルと仮定すると、ここから視聴履歴に基づいて特定の番組を視聴したセグメントを抽出する場合、世帯視聴率[2]が 10%の番組だとセグメントサイズは 200 程度まで減少する。このセグメントの視聴傾向を分析するために別の番組の世帯視聴率を求めようとすると、真の世帯視聴率が 10%の番組の場合、表 2 より±4.2%の標本誤差を含むことが分かる。

一方で、近年、ネットワーク接続型テレビの普及により、視聴者から利用許諾を得て取得した視聴データを用いて番組やテレビ広告の視聴分析が可能となりつつある。テレビ視聴データ集計システム TimeOn Analytics では、2018 年 6 月現在、全国64万台の機器から利用許諾を得て視聴データの提供を受けている [5] [6]。各機器について秒単位のリアルタイム視聴時間とタイムシフト視聴時間を集計できる。本システムの大規模テレビ視聴データを用いれば、先述の例のようにセグメント抽出によってサンプル数が数百分の一となっても、精度の高い分析結果が得られる。

そこで本稿では、TimeOn Analytics のテレビ視聴データを用い、広告効果の高い番宣出稿のための、行動学的属性に基づくセグメント抽出手法を提案する。

表 1　性別・年齢区分の例

区分	性別	年齢
T	-	～19 歳
M1	男性	20 歳～34 歳
M2	男性	35 歳～49 歳
M3	男性	50 歳～
F1	女性	20 歳～34 歳
F2	女性	35 歳～49 歳
F3	女性	50 歳～

表 2　標本誤差早見表（信頼度 95%）

		サンプル数：n				
母比率：p		20	200	2000	20000	200000
5%	95%	±4.4%	±1.4%	±0.4%	±0.1%	±0.0%
10%	90%	13.4	4.2	1.3	0.4	0.1
20%	80%	17.9	5.7	1.8	0.6	0.2
30%	70%	20.5	6.5	2.1	0.7	0.2
40%	60%	21.9	7.0	2.2	0.7	0.2
50%	50%	22.4	7.1	2.2	0.7	0.2

（注）[2]を参考にサンプル数を変えて標本誤差を算出。

2. 提案手法

2.1 セグメント抽出手法

番宣のターゲットは宣伝を行う番組に興味を持つ視聴者である。そこで本稿では、視聴者が過去に視聴した番組がその視聴者の関心を反映していると考え、当該コンテンツと関連性が高い番組を過去に視聴していた視聴者をターゲットとする。例えば、あるドラマシリーズ第 1 期が終わった後に第 2 期の番宣を新たに出稿する場合、過去に第 1 期の番組を視聴していた視聴者は、番宣接触によって第 2 期の番組の視聴を促しやすい有効ターゲットとなりうる。他も、同じ出演者、同じ番組ジャンル、同じ放送時間帯など、様々な切り口から過去番組を設定することができる。従来の人口統計学的属性に基づく手法（従来手法）では性別・年齢などによって視聴者の趣味趣向が類似性すると仮定していたのに対し、提案手法では過去番組の視聴実績の有無でそれを仮定している。そのため、コンテンツ制作者が想定していなかったターゲットからも支持されるようなコンテンツの場合、従来手法では抽出できなかった有効ターゲットについても提案手法により抽出可能になる。例えば、

†東芝映像ソリューション株式会社
R&D センター クラウド技術開発部

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

図 1 に示すように、主婦向けに制作されたドラマ番組 A の第 2 期放送回の番宣を出稿するケースを考えてみる。従来手法では、番組 A が主婦向けに制作されたことを考慮し、20～35 歳の女性の視聴者を番宣のターゲットセグメントと設定する。しかし、番組制作者の期待とは裏腹に、番組 A が 20～35 歳の男性にも支持されるような内容であった場合、従来手法では、20～35 歳男性という潜在的な有効ターゲットがセグメントに含まれないという問題がある。一方、提案手法では、ドラマ A の第 1 回放送回の視聴者を番宣のターゲットセグメントとする。第 1 回放送回の視聴者の中には、20～35 歳男性の視聴者が一定のボリュームで含まれると期待されるため、従来手法では見逃していた有効ターゲットもセグメントに含めることができる。

本稿では問題の簡略化のため、番組p_mを宣伝するための番宣の出稿に際して前回放送回の番組p'_mを関連性が高い番組とした場合について述べることとする。提案手法のセグメント抽出の手順を以下に示す。

1. 視聴データ提供機器d_nそれぞれについて、p'_mのリアルタイム視聴秒数とタイムシフト視聴秒数の合計値$t'_{m,n}$を算出する。
2. ザッピング視聴等の影響を排除するために閾値α（$0 < \alpha < 1$）を設け、p'_mの放送秒数t'_mに対し、$t'_{m,t} > \alpha t'_m$を満たす機器d_nを抽出する。
3. 2.で抽出された機器集合$\boldsymbol{S}'_m$を、p_mの番宣出稿におけるターゲットセグメントとする。

2.2 テレビ広告出稿手法

セグメント$\boldsymbol{S}'_m$を抽出した後は、ターゲットがテレビをよく視聴している時間帯を明らかにすることで、テレビ広告を出稿するべきタイミングの検討の一助とする。一般に番組の内容は曜日によって決まっていることが多いため、人のテレビ視聴傾向も曜日単位で異なって現れる。そこで本稿では、「月曜日 11:00～12:00」のように、曜日と 1 時間おきの時間帯で定義される各時間枠について、ターゲットの平均的なテレビ視聴時間を算出する。2.1 節ではリアルタイム視聴時間とタイムシフト視聴時間を合計して計算したが、タイムシフト視聴においてはテレビ広告がスキップされやすい傾向があるため、ここではリアルタイム視聴時間のみを扱うこととする。その後、各時間枠を曜日-時間帯軸上にプロットし、平均視聴時間の長さにしたがって色付けした視聴ヒートマップを作成する。視聴ヒートマップの例を図 2 に示す。図 2 では平均視聴時間が長いほど濃い色で表示している。図 2 を見て分かるように、視聴量をヒートマップ化することで、ターゲットの視聴がどのように分布しているかを一覧できる。ターゲットの視聴が多い時間枠に番宣出稿すれば、効率良く番宣出稿できる。

提案手法の番宣出稿の手順を以下に示す。

1. 対象チャンネルにおける各時間枠w_l（$l = 24$ 時間×7 日間）について、ターゲット機器$d_n \in \boldsymbol{S}'_m$のリアルタイム視聴時間の平均値を算出する。
2. 算出された値を曜日-時間帯軸上でヒートマップ化する。
3. ターゲットの視聴量が特に多い枠から順に広告を出稿する枠の候補とし、広告出稿に関する他の制約条件を考慮しながら出稿する時間帯を決定する。

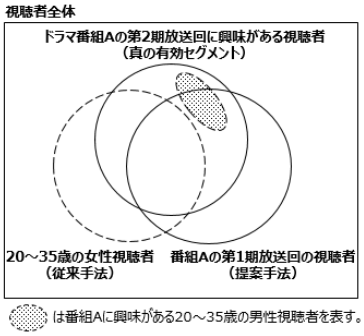

図 1　従来手法と比較手法のアプローチの例

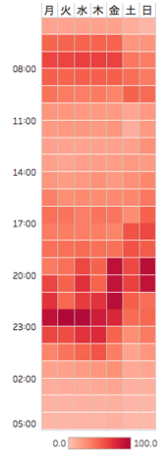

図 2　視聴ヒートマップの例（色が濃いほど多く視聴されていることを示す）

3. 実験方法

提案手法の有効性を実験により検証する。実験は 2 種類行い、実験 1 では真に有効な機器がセグメントにどの程度含まれるかを定量的に評価する。実験 2 では抽出セグメントの視聴傾向が真に有効なセグメントとどの程度異なるかについて典型例を取り上げて定性的に評価する。

3.1 実験条件

はじめに両実験における実験条件について述べる。関東地区における視聴データ提供機器を対象とし、2018 年 2 月 8 日 5 時から 2018 年 2 月 9 日 5 時の期間において関東キー局で放送された各番組p_mについて、番宣出稿のターゲットとなるセグメント$\boldsymbol{S}'_m$を抽出する。セグメント抽出に用いる過去番組p'_mは、一週間前の 2018 年 2 月 1 日 5 時から 2018 年 2 月 2 日 5 時に放送された番組のうち、当該番組と番組開始時間・終了時間が一致する番組とする。一致する番組が存在しない場合は今回の実験の対象外とする。なお、p'_mの視聴判定条件はα=0.5、つまりリアルタイム視聴時間とタイムシフト視聴時間の合計が番組放送時間の 50%以上であることとする。実験条件を表 3 にまとめる。

表 3　実験条件

地域	関東
対象機器数	n=229,014
放送局	関東キー局 5 局
比較対象番組	「2018 年 2 月 1 日 5 時～2018 年 2 月 2 日 5 時」の期間に放送された番組
評価対象番組	「2018 年 2 月 8 日 5 時～2018 年 2 月 9 日 5 時」の期間に放送された番組
対象番組数	m=241
視聴判定条件	α=0.5

3.2 実験 1

実験 1 では提案手法で得られたセグメント$\boldsymbol{S}'_m$と真の有効セグメント$\boldsymbol{S}_m$の重なりを定量的に評価する。$\boldsymbol{S}_m$を観測することは難しいため、本実験では、当該番組p_mを実際に視聴した機器の集合を$\boldsymbol{S}_m$の Ground Truth とする。そして、$\boldsymbol{S}'_m$と$\boldsymbol{S}_m$の視聴機器の重なりを再現率（recall）、適合率

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

（precision）、F 値（F-measure）の 3 つの評価値を用いて評価する。各評価値の関係を図 3 に示す。再現率は、正と予測したデータのうち、実際に正であるものの割合を示す。図 3 中の記号を用いると、TP/(TP+FN)で与えられる。適合率は、実際に正であるデータのうち、正であると予測されたものの割合を示す。同様に、TP/(TP+FP)で与えられる。F 値は再現率と適合率の調和平均である。つまり、今回の番宣出稿の問題において、再現率が高いほど$\boldsymbol{S'}_m$に含まれる真に有効な視聴者は多く、適合率が高いほど$\boldsymbol{S'}_m$に含まれる冗長な視聴者は少ないことを意味する。そして、F 値はその両方を評価する評価値である。

比較手法として、性別・年齢を用いてセグメントを作成した場合についても考える。TimeOn Analytics の視聴データ提供機器の一部には、性別・年齢を含むデモグラフィック属性が付属しているため、これを利用する。セグメント作成には、表 4 で定義される区分を用いる。区分の全てを組合せた2^{10}通りのパターンを列挙してそれぞれに対応する機器集合を求め、番組毎に$\boldsymbol{S}_m$に最も当てはまりの良い機器集合を採用し、これを比較手法における抽出セグメント$\boldsymbol{S''}_m$とする。当てはまりの良さの評価値としては F 値を用いる。なお、デモグラフィック属性はアンケート（任意回答）に回答した機器にしか付属していないため、$\boldsymbol{S''}_m$を作成・評価する際には、$\boldsymbol{S}_m$に含まれる機器のうち、アンケートに回答していない機器を事前に取り除く。その後、再現率、適合率、F 値が提案手法と比較手法でどの程度異なるかを確認する。

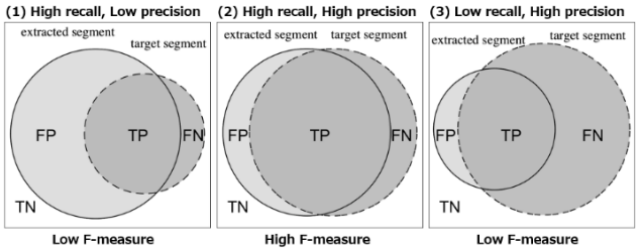

図 3　再現率と適合率と F 値の関係

表 4　本実験における性別・年齢の区分

区分	性別	年齢
C	-	～12 歳
T	-	13 歳～19 歳
M1	男性	20 歳～34 歳
M2	男性	35 歳～49 歳
M3	男性	50 歳～64 歳
M4	男性	65 歳～
F1	女性	20 歳～34 歳
F2	女性	35 歳～49 歳
F3	女性	50 歳～64 歳
F4	女性	65 歳～

3.3 実験 2

実験 2 では、実験 1 で得られた各セグメント$\boldsymbol{S}_m$、$\boldsymbol{S'}_m$、$\boldsymbol{S''}_m$の典型例について、3.2 節で述べた手順で視聴ヒートマップを作成する。その後、各ヒートマップを比較することで、それぞれのセグメントの視聴傾向がどの程度類似しているかを定性的に評価する。なお、ヒートマップは過去番組p'_mが放送された週である 2018 年 1 月 29 日 5 時から 2018 年 2 月 5 日 5 時までの 1 週間を対象として作成する。

4. 実験結果

4.1 実験 1

実験で得られた各セグメント$\boldsymbol{S'}_m$、$\boldsymbol{S''}_m$について、再現率、適合率、F 値の番組平均を表 5 に示す。また、再現率、適合率の散布図を図 4 に示す。表 5 を見ると、全ての評価値について、提案手法では比較手法よりも平均して 10 倍以上高い値を示している。また、図 4 を見ると、比較手法における適合率と再現率の分布の仕方は異なっており、適合率が比較的高いものの再現率は低い番組が多く存在していることが分かる。一般に、セグメントに含まれる有効でない機器の割合を減らすため、条件を絞って抽出機器を減らすほど、正確性を示す適合率は増加するが、その分有効なターゲット機器も除外される可能性があるため、網羅性を示す再現率は減少する傾向にある。この傾向は図 3 における（3）の状態と一致し、真に有効な視聴者を十分カバーできていないことを示す。例えば、子供向けの番組の番宣のターゲットを設定するにあたり、子供の視聴者は有効なターゲットとして抽出されているが、子供以外の有効なターゲットを多く取りこぼしてしまう状態がこれに相当する。一方、提案手法では、再現率と適合率のバランスが取れたセグメントが抽出されており、有効なターゲットが極端に少ないセグメント、あるいは有効でないターゲットが極端に多く含まれるセグメントが抽出されにくい手法であると言える。

表 5　各評価値の番組平均

	再現率平均	適合率平均	F 値平均
提案手法	0.31	0.29	0.30
比較手法	0.02	0.07	0.03

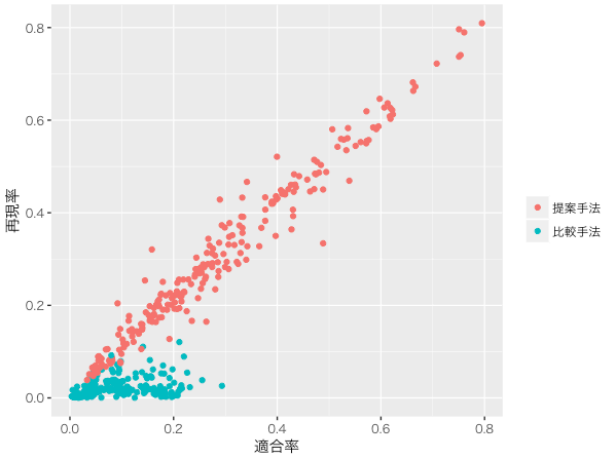

図 4　再現率、適合率の散布図

4.2 実験 2

次に、各セグメント$\boldsymbol{S}_m$、$\boldsymbol{S'}_m$、$\boldsymbol{S''}_m$について、視聴傾向の差異を定性的に評価する。ここでは、比較手法と提案手法のそれぞれにおいて、F 値が特に高かった 2 番組を取り上げる。図 5 は比較手法の F 値が 0.15 であった放送局 A のバラエティ番組 a について、図 6 は提案手法の F 値が 0.74 であった放送局 B のアニメ番組 b について、それぞれ各セ

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

グメントの視聴ヒートマップを比較したものである。なお、ヒートマップごとに平均視聴時間の値域は異なるため、平均視聴時間の最大値が 100 となるようにそれぞれ正規化を行っている。

図 5 おける正解セグメントの視聴ヒートマップを見ると、「木曜 23 時台」の時間枠が最も視聴されているが、この枠は番組 a が放送されている時間帯である。次に視聴されているのが「木曜 24 時台」、「木曜 22 時台」、「日曜 19 時台」、「日曜 20 時台」の時間枠であり、この時間帯に番組 a の番宣を出稿することでターゲットに効率よく接触させることができることを示す。ここで、比較手法で抽出されたセグメントの視聴ヒートマップを見ると、視聴の多さを示す色の濃さは異なるものの、先述の時間枠がよく視聴されており、正解セグメントの視聴傾向を再現できている。この番組は、若者だけに多く視聴される傾向があるため、従来手法でも真のターゲットセグメントに近いセグメントを抽出できたと考えられる。一方、提案手法で抽出されたセグメントも、同じ視聴傾向にあることが分かる。

図 6 を見ると、正解セグメントは「木曜 19 時台」の時間枠を最も多く視聴しており、この枠はやはり番組 b が放送されている時間帯であった。次に多く視聴されているのは「日曜 8 時台」、「木曜 18 時台」、「金曜 18 時台」である。提案手法で抽出されたセグメントについても同じ枠が多く視聴されており、正解セグメントと視聴傾向がほぼ一致している。このうち、「日曜 8 時台」に放送されている番組を調査したところ、番組 b でも放送されていたアニメ作品に関する情報バラエティ番組であった。抽出されたセグメントは同作品に関心があると予想されるため、同作品に関する番宣を出稿する際の有効なターゲットとなり得る。一方、比較手法で抽出されたセグメントは、全く異なる視聴傾向をしていることが分かる。この番組は、複数の性別・年齢区分をターゲットに含むため、真のターゲットセグメントに近いセグメントを抽出できなかったと考えられる。

5. おわりに

本稿では、視聴者の行動履歴である大規模テレビ視聴データを用い、行動学的属性に基づくセグメント抽出手法を提案した。提案手法は、特定の番組に興味を持つ視聴者を抽出する実験において、人口統計学的性年代属性に基づく手法よりも、F 値が 10 倍以上高いセグメントを抽出できることがわかった。また、抽出されたセグメントのテレビ視聴時間を時間帯毎に可視化することで、ターゲットの特徴的な視聴傾向を確認することができた。

提案手法によると広告効果が高い時間枠の候補が複数得られるため、広告出稿に関する他の制約条件を考慮しながら、柔軟で広告効果の高い広告出稿が可能になる。本手法は、テレビメディアにおけるターゲティング広告への活用が期待される。

今後は、本稿では扱わなかった広告出稿の効果検証などについて、検討を進めていく。

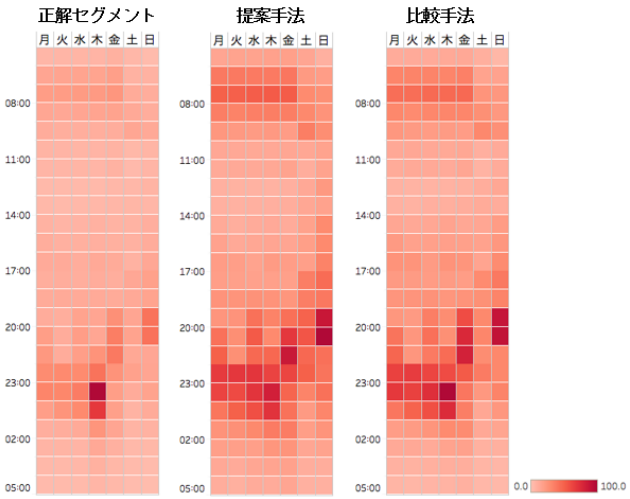

図 5　視聴量ヒートマップの比較（A 局バラエティ番組 a）

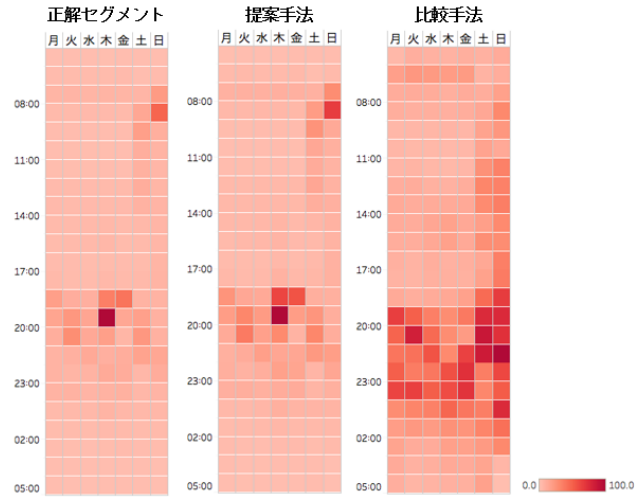

図 6　視聴量ヒートマップの比較（B 局アニメ番組 b）

参考文献

[1] 株式会社 電通, “2017 年　日本の広告費”, http://www.dentsu.co.jp/news/release/2018/0222-009476.html.
[2] 株式会社 ビデオリサーチ, “視聴率ハンドブック”, p.4,「世帯視聴率」と「個人視聴率」, https://www.videor.co.jp/tvrating/pdf/handbook.pdf.
[3] 総務省 情報通信政策研究所, “行動ターゲティング広告の経済効果と利用保護に関する調査研究報告書”, 2010.
[4] 株式会社 スイッチ・メディア・ラボ, “テレビ視聴分析を、もっと気軽に、もっと簡単に”, https://www.switch-m.com/smart/.
[5] 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, 2016.
[6] 東芝映像ソリューション株式会社, “東芝テレビ視聴データの特長”, http://www.toshiba.co.jp/tvs/tvdata/feature/index_j.htm.
[7] 東芝映像ソリューション株式会社, “レグザクラウドサービス「TimeOn」”, http://m.timeon.jp/
[8] 株式会社 ビデオリサーチ, “[視聴率データ]週間高世帯視聴率番組１０．”, https://www.videor.co.jp/tvrating/.

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.