H-078

# 電子番組表に基づいた番組紹介映像の自動生成手法の評価

# Evaluation of Automated Production Method for TV Program Trailer using Electronic Program Guide

河合 吉彦 †　　住吉 英樹 †　　八木 伸行 †
Yoshihiko Kawai　Hideki Sumiyoshi　Nobuyuki Yagi

## 1　まえがき

大量の映像から，目的の映像を効率的に探索するための技術のひとつとして映像の要約がある．我々は，要約映像の一種である番組紹介映像を対象に研究を進めており，電子番組表（EPG:Electronic Program Guide）を利用した自動生成手法を提案した [1]．本研究における課題のひとつは生成された紹介映像の評価方法である．これまでは実際に放送された番組スポットとの比較によって定量的な評価を試みていた．しかし，放送スポットが絶対的な正解データとは言えず，これを基準とした評価方法には問題があった．そこで，本稿では同一の番組に対して，人手による正解データを複数用意し，それらと比較することによって，より客観的な評価を試みる．

## 2　EPG を利用した番組紹介映像の生成

提案手法 [1] の内容を簡単に説明する．本手法は，対象番組の EPG の有無によって次の 2 つの手法からなる。

### 2.1　手法 1:類似度に基づく手法

EPG が入手できる場合は，EPG に記載される紹介テキスト（EPG テキスト）と CC の類似度に基づいて紹介映像を生成する．図 1 に手法の概要を示す．まず EPG テキストの各文に対して，最も類似度の高い文を CC から抽出する．次に，抽出した CC 文に対応する映像区間を連結し番組紹介映像とする．

文の類似度は，共通に含まれる語に基づいて算出する．このとき，助詞などの一般的な語よりも，固有名詞など希少性の高い語が共通に含まれるほど類似度が高くなるよう考慮する．また，CC の一部にのみ出現するような語が共通に含まれるほど類似度が高くなるよう考慮する．提案手法では，ベイズ信頼度ネットワークを利用して文の類似度を定義する．

信頼度ネットワークは，番組 CC の各文を表すノード $d_i(i = 1,..,D)$ と，索引語を表すノード $t_j(j = 1,..,T)$，EPG テキストの一文を表す $q$ からなる [2]．また，ノード $t_j$ は，0（CC 文に含まれない）または 1（含まれる）の状態を持つ．類似度の算出式を式（1）に示す．

$$P(q|d_i)=\sum_{all\ s_k} P(q|t_{1s_1},..,t_{Ts_T})\prod_{j=1}^{T} P(t_{js_j}|d_i) \qquad (1)$$

$P(t_{js_j}|d_i)$ は，番組 CC における索引語の出現頻度から算出する．また，$P(q|t_{1s_1},..,t_{Ts_T})$ は，索引語の希少性が高いほど，値が大きくなると仮定する．具体的には，過去の放送番組におけるエントロピに基づいて定義する．

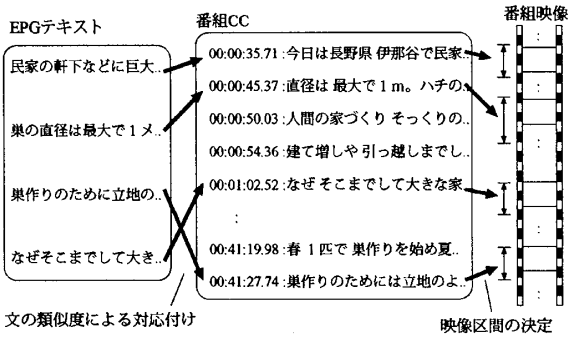


図 1　手法 1 の概要

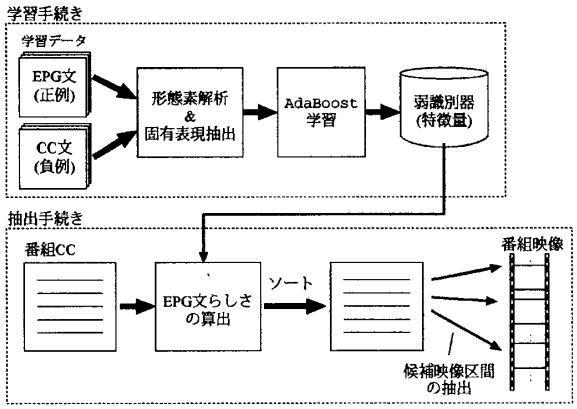


図 2　手法 2 の概要

### 2.2　手法 2:類似度に基づく手法

EPG の入手が困難な番組に対しては，一般的な EPG テキストが持つ文章特徴に基づいて紹介映像を生成する．手法の概要を図 2 に示す．学習手続きでは，学習用に収集した EPG 文と CC 文から AdaBoost によって識別器を学習する．抽出手続きでは，識別器によって CC の各文がどの程度 EPG 文らしいかを算出し，値が高いものから順にある文数だけ選択する．最後に，選択された CC 文に対応する映像区間を連結し紹介映像とする．

学習データには，正例として EPG 文を，負例として番組 CC 文を使用する．AdaBoost 学習には次の特徴量を使用する．これらは，特徴抽出の容易性と学習処理における計算負荷を考慮して選択した．

(a) 形態素数が閾値以上，もしくは閾値以下
(b) ある品詞が含まれる，もしくは含まれない
(c) ある索引語が含まれる，もしくは含まれない
(d) ある固有表現が含まれる，もしくは含まれない

固有表現は組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現の 8 種類を使用する [3].

† NHK 放送技術研究所

表 1　正解データとの比較結果（10 番組分）

| | 正解データ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 放送スポット | | 専門家 1 | | 専門家 2 | | 合計 | |
| | 再現率 | 適合率 | 再現率 | 適合率 | 再現率 | 適合率 | 再現率 | 適合率 |
| 放送スポット | – | – | 0.53 | 0.44 | 0.41 | 0.41 | 0.47 | 0.43 |
| 専門家 1 | 0.44 | 0.53 | – | – | 0.38 | 0.42 | 0.41 | 0.47 |
| 専門家 2 | 0.41 | 0.40 | 0.42 | 0.38 | – | – | 0.41 | 0.39 |
| 手法 1 | 0.40 | 0.45 | 0.39 | 0.40 | 0.36 | 0.44 | 0.39 | 0.43 |
| 手法 2 | 0.23 | 0.28 | 0.15 | 0.22 | 0.19 | 0.25 | 0.19 | 0.25 |
| 固定間隔 | 0.10 | 0.09 | 0.14 | 0.13 | 0.09 | 0.11 | 0.11 | 0.11 |

## 3　評価実験

提案手法を用いて，実際に放送された自然番組，10 番組を対象に紹介映像を自動生成した．実験の結果，手法 1，2 とも一番組あたり 43 分の本編映像に対して約 30 秒の紹介映像を自動生成することができた．

生成された番組紹介映像を定量評価するために，絶対的な正解データを設定することは困難である．紹介映像の内容は，何をどのように紹介するかといった制作者の感性や主観によって変化するためである．そこで，本実験では同一の番組に対して，人手による正解データを複数用意し，それらと映像内容を比較する事によって，より客観的な評価を試みる．さらに，人手による正解データ間の重なりを調査することにより，精度の上限を推定することも可能になると考える．具体的な正解データとしては，実際に放送された 30 秒のスポット映像，および編集業務に携わった経験のある専門家による紹介映像を用いた．専門家による正解データは，専門家 2 名に対して，EPG テキストは提示せず番組本編映像のみを視聴させ紹介映像に適したシーンをそれぞれ列挙してもらうことにより作成した．このとき映像長は 30 秒程度になるよう依頼した．映像の比較はショット単位で実施し，評価には次式で表される再現率および適合率を用いた．

$$\text{再現率} = \frac{N_b}{N_g}, \qquad \text{適合率} = \frac{N_b}{N_o} \tag{2}$$

$N_g$ は正解データに含まれていたショット数，$N_o$ は提案手法によって抽出されたショット数を表す．また $N_b$ は，正解データに含まれるショットのうち，提案手法でも抽出できた数を表す．なお，本実験では映像内容が同一であれば，画角などが多少異なっても一致と判定した．

精度の下限を調査するため，番組内容を考慮せず機械的に映像を短縮するような手法を用いた．具体的には，番組冒頭から 90 秒の位置を始点に 10 分間隔で 6 秒ずつの部分映像を取り出し，合計で 30 秒の映像を生成するような手法である．ここで開始点を 90 秒としたのは，冒頭のオープニング部分を除いた番組の本編部分のみを対象とするためである．

正解データを相互に比較した結果を表 1 の上部に示す．実験の結果，それぞれ正解データに含まれるショットは，再現率，適合率が約 40%〜55% という精度で映像内容が一致していることが分かった．人手によって制作された正解データ間の重なりがこの程度であることから，自動生成における精度の上限も 50% 前後であると考えられる．次に，表 1 の下部に提案手法による紹介映像と正解データとの比較結果を示す．手法 1 については平均で再現率 39%，適合率 43% という結果が得られた．正解データ間の比較結果を考慮すると，非常に良好な結果が得られたといえる．この結果から，対象番組の EPG テキストに基づいてシーンを選択する手法 1 の有効性が確認できた．それに対して，手法 2 は平均で再現率 19%，適合率 25% という結果となり，手法 1 に比べて精度が低下した．手法 2 は対象番組に対する知識がない状態で映像区間を選択しているため，精度の低下はやむを得ないと考える．しかしながら，固定間隔でサンプリングする単純な手法が平均再現率 11%，平均適合率 11% であることから，一般的な EPG 特徴のみでも，ある程度は番組の意味内容を考慮したシーンの選択ができたと言える．手法 1，2 のいずれも，適合率に比べて再現率が低い値となっている．今後，映像や音声情報も統合的に扱うことにより，CC のみでは抽出の難しいシーンも考慮できるよう検討が必要と考える．

## 4　あとがき

本稿では，電子番組表に基づいて番組紹介映像を自動生成する 2 つの手法に対して，手法の有効性を客観的に評価することを試みた．評価実験では，同一の番組に対して複数の正解データを用意し，それらと映像内容を比較することにより定量的な評価を試みた．実験の結果，対象番組の EPG に基づく手法は，再現率 39%，適合率 43% という結果が得られた．また，一般的な EPG 特徴に基づく手法は，再現率 19%，適合率 25% という結果が得られた．機械的に映像を短縮する手法と比較して，いずれの手法も高い精度となっており，番組内容を考慮したシーン選択が実現できていることが確認できた．今後は，シーンの順序やつながりなども考慮した評価手法の検討や，主観評価実験なども検討したい．また，今回の実験結果から，映像情報や音声情報などを統合的に利用するなど，手法の改良についても検討したい．

## 参考文献

[1] 河合吉彦, 住吉英樹, 八木伸行, “電子番組表を利用した番組紹介映像の自動生成手法,” 信学技報, vol.106, no.543, pp.165–170, 2007

[2] H.Turtle and W.B.Croft, “Evaluation of an interface network-based retrieval model,” ACM Trans. Information Systems, vol.9, no.3, pp.187-222, 1991.

[3] Information Retrieval and Extraction Exercise (IREX), ”NE ルール, 定義（バージョン 990214）,” http://nlp.cs.nyu.edu/irex/NE/