Vol.2010-HCI-137 No.8
2010/3/19

テレビ番組間の関連性に基づくコンテンツ検索システム「ローミングナビ」の開発

三原 功雄[†] 関根 真弘[†] 樋口 靖和[†]
高倉 潤也[†] 鈴木 優[††] 山内 康晋[†]

本論文では，我々が新規開発したテレビ番組間の関連性に基づくコンテンツ検索システム「ローミングナビ」に関して論じる．ローミングナビは，膨大な映像コンテンツの中から，自分の興味にあったコンテンツを効率良く探すことができるシステムである．ユーザーの感じるテレビ番組間の関連性を，被験者実験およびコンジョイント分析，項目応答理論を始め様々な統計手法を用いて高い精度（適合率 96.9%，ユーザーカバー率 80.8%）で再現していることが特徴である．ローミングナビにより，“関連”を頼りに関心のあるコンテンツを渡り歩く新しい視聴スタイルが可能となった．

Development of “Roaming Navi”: The content finding navigation system based on the relevance between TV program contents

Isao Mihara[†] Masahiro Sekine[†] Yasukazu Higuchi[†]
Junya Takakura[†] Masaru Suzuki[††] and Yasunobu Yamauchi[†]

Users can become overwhelmed by the vast amounts of visual contents available from various sources including digital terrestrial and satellite broadcasting, IP TV, and much more. We have developed the content finding navigation system called “Roaming Navi” that allows the user to swiftly discover the types of content corresponding to his or her interests in this ocean of visual contents. Our relevance calculation method is based on user sensibility model that is availed by statistical method and has high reliability (precision 96.9%, user cover rate 80.8%).

1. はじめに

近年，テレビ放送の多チャンネル化，IP テレビやインターネット動画配信サービスの開始，大容量の記憶装置を備えたデジタル機器の普及などにより，我々の身近には膨大な映像コンテンツが溢れている．一方，日常生活の中でのテレビ離れが進んでいると言われている．その理由として，「他にやることの増加によるテレビを見る時間の減少」，「見たい・面白い番組がない」が挙げられている [1]．近年の生活スタイルの変化により，他の行動に比べてテレビ視聴の優先順位が低下しているということである．視聴可能な映像コンテンツが増加しているにも関わらず見る時間がない．さらに，膨大な映像の中に見たいものがあるかも知れないにも関わらず，それを知る術がないため，見たい・面白い番組が無いと思いこんでいるのではないだろうか．このような状況のもとでは，膨大なコンテンツの中から効率的にユーザーの興味のあるコンテンツを探し，限られた時間の中で如何に効率的に呈示するかが求められている．

そこで我々は，膨大な映像コンテンツの中からユーザーの興味にあったコンテンツを効率良く探すことができるシステムとして，テレビ番組間の関連性に基づくコンテンツ検索システム「ローミングナビ」を開発した [4], [6], [7].

2. ローミングナビとは？

図 1 に「ローミングナビ」の初期画面（メイン画面）を示す．ローミングナビは，テレビ番組間の関連性に基づくコンテンツの検索システムである[a].

ユーザーが注目しているテレビ番組コンテンツ（以降，注目コンテンツ）が画面中央に配置され，これと関連するコンテンツ（以降，関連コンテンツ）群がこの周辺に関連性の強さに応じて同心楕円状に配置される（図 2）．注目コンテンツに近い位置ほど関連性が強く，遠くなるほど関連性が弱い関連コンテンツが配置される．

画面は分類軸により上下左右の 4 つのエリアに分かれており，関連コンテンツは関連性の根拠によって分類されて配置される．分類軸は，タイトル（上方向），人物（右方向），キーワード（下方向），ジャンル（左方向）である．

ユーザーは，リモコンやマウスを用いて，配置された関連コンテンツの間をカーソル移動することができる（図 1 では現在フォーカスが当たっている関連コンテンツが拡大表示されている）．

[†]株式会社 東芝 研究開発センター ヒューマンセントリックラボラトリー
Humancentric Laboratory, Corporate Research & Development Center, TOSHIBA Corporation
[††]株式会社 東芝 研究開発センター 知識メディアラボラトリー
Knowledge Media Laboratory, Corporate Research & Development Center, TOSHIBA Corporation

a) ローミングナビが検索対象とするテレビ番組は，ユーザーの視聴時に今後放送予定の番組（未来の番組で EPG 画面に存在する番組）およびユーザーが録画した番組である．ローミングナビでは録画済みテレビ番組コンテンツは図 1 のようにサムネイル付きで表示し，放送予定番組は図 3 のように番組内容が表示される．

Vol.2010-HCI-137 No.8
2010/3/19

図 1　ローミングナビのＧＵＩ（メイン画面）

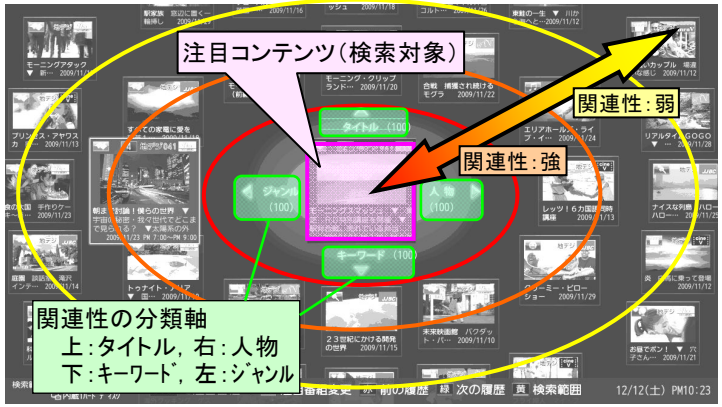


図 2　画面の説明（メイン画面）

ユーザーのカーソル移動動作により，フォーカスが現在の画面範囲内を超えた場合，現在フォーカスが当たっている関連コンテンツが存在する分類軸方向に画面がスクロールし，特定軸内での関連コンテンツの配置画面に遷移する．図 3 は人物軸方向（メイン画面の右方向）へカーソルを進めた場合の例である．人物軸に配置された関連コンテンツが，メイン画面と同様に関連性の強さに応じて配置されている．さらにカーソルを進める（図 3 の場合は，さらに右方向へ進める）と，この画面内で注目コンテンツの表示を残したまま，関連コンテンツ群の呈示部分がスクロールされ，より下位の関連コンテンツにアクセスすることが可能となる．

関連コンテンツにフォーカスが当たると，その関連コンテンツにおける注目コンテンツとの関連情報（関連性の根拠）がポップアップ表示される．図 3 にて，フォーカスされている関連コンテンツの左横にポップアップウィンドウが出ている様子が確認できる．関連情報としては，番組タイトルの類似具合，注目と関連コンテンツで共に出演している人物や共に扱っている話題を表すキーワード，共通しているジャンルなどが表示される．

図 3　人物軸方向（右方向）へカーソルを進めた場合の画面

現在フォーカスされている関連コンテンツを選択すると，この関連コンテンツを新たな注目コンテンツとして再検索が行われ，画面の再構築が行われる．ユーザーはこの動作を繰り返しながら，自分の興味のあるテレビ番組を発見することが可能となる．

ローミングナビは，従来のキーワード入力型のテレビ番組検索システムと異なり，キーワードを入力することなく，番組間の関連性に基づいて配置・呈示された関連コンテンツを辿ることを繰り返しながら，様々なコンテンツにアクセス可能となるため，今まで自分が見たこともない番組，思いがけないキーワードで繋がっていた番組などを探すことが可能となり，新たな番組の発見を創出することができる新感覚のコンテンツ検索システムであると言える．

3. テレビ番組間の関連性の算出

3.1 算出処理フロー

図 4 にローミングナビにおけるテレビ番組間の関連性の算出処理フローを示す．処理は大きく分けて以下の 4 段階から構成されている．

(1) 項目要素抽出処理
(2) 類似度算出処理
(3) 関連度算出処理
(4) レイアウト算出処理

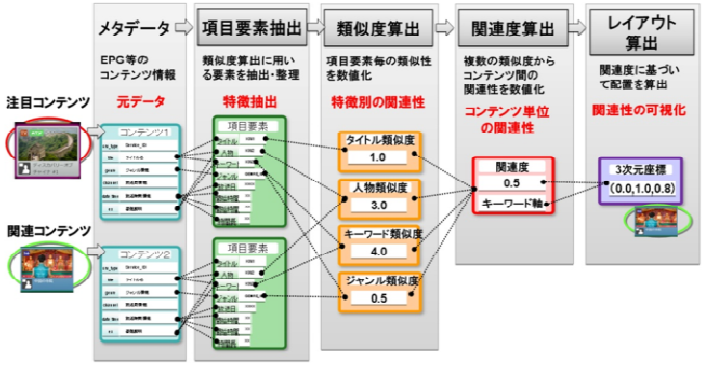


図 4　関連性算出の処理フロー

3.2 項目要素抽出処理

項目要素抽出処理は，テレビ番組に付随しているメタデータ（番組情報が記されている）から情報を抽出し，タイトル要素（番組名を構成している語句群），人物要素（番組情報に含まれている人名群），キーワード要素（番組情報に含まれている人名以外のキーワード群），ジャンル要素（番組情報に含まれているジャンル情報群），放送時間要素（番組情報に含まれる放送開始・終了時間などの情報），放送局要素（番組情報に含まれる放送局情報）の 6 つの要素（以降，項目要素と呼ぶ）別に再構成する処理である．この際，該当項目要素として取得された語句がメタデータの何に由来して抽出されたかの情報（抽出元情報）も同時に記録しておく．例えば，人名項目要素は A, B から成り，A は番組概要から，B は番組名から抽出された，というようにである．

テレビ番組のメタデータには，番組名，番組概要，番組内容詳細などが記されているが，これらは非定型の自然言語のテキストによって提供され，その内容も記載方法も放送局により様々である．そこで，メタデータから人名，キーワードなどを取得するためには，テキスト文中からこれらを抽出する技術が必要となる．今回，メタデータからのキーワード（人名含む）の抽出は，当社が独自に技術開発を行ってきた語句意味解析技術[2], [3] を利用した．語句意味解析技術は，テキスト文を解析し，そこに含まれるキーワード（人名を含む）を，その語句の意味カテゴリと併せて抽出する技術である．これを用いることで，著名人名，政治家名，歴史上の人物，キャラクター名，地名，組織名，スポーツ用語，健康・医療用語，…，などといった約 100 種類の意味カテゴリとともに語句の抽出を行うことが可能となる．

3.3 類似度算出処理

類似度算出処理では，項目要素抽出処理によって抽出された各項目要素毎に類似度を算出する．類似度は，基本的には，その項目要素に含まれる語句の一致度合いであり，2 つのテレビ番組を項目要素毎で比較した場合の個々の類似性を示している．

例えば，テレビ番組 A の人物要素が X, Y, W であり，テレビ番組 B の人物要素が X, Z, W の場合，この 2 つの番組に共通している人物は X, W であるから，人物要素における類似度 $S_{person}(A,B) = 2.0$ となる．

この際，項目要素抽出処理の際に得られた各要素に付随している語句意味カテゴリ，抽出元情報を加味して類似度の重み付けを行うこともある．抽出元情報の場合，例えば，番組名中に人名が現れている場合には，その人の冠タイトルの番組であるため，その抽出人名の重みを高くして優位に扱うといったことが考えられるし，付随している語句意味カテゴリが具体的・特徴的な意味を持っている場合（例えば，特定の商品名などの場合）には，同様に重みを高くした方が良い場合が多い．

同様に，タイトル要素の類似度 S_{title},キーワード要素の類似度 $S_{keyword}$,ジャンル要素の類似度 S_{genre},放送時間要素の類似度 S_{time},放送局要素の類似度 $S_{broadcaster}$ に 関しても算出する．

3.4 関連度算出処理

関連度算出処理では，類似度算出処理によって得られた項目要素毎の類似度を総合的に加味し，以下に示す式(1)を用いて 2 つのテレビ番組間の関連度 R を算出する．

$$R = f(S_{title}, S_{person}, S_{keyword}, S_{genre}, S_{time}, S_{broadcaster}) \quad (1)$$

ここで，関数 f は，2 つの番組の関連性が高いほど高い関連度の値が算出される関数である．

2 つのテレビ番組間の関連性，つまり，式(1)における関数 f とはどういったものであろうか？前節で求めた各々の項目要素における類似度がどのような状態になった場合に，人は番組間の関連性が高いと感じるのだろうか？また，その感じ方は共通の感じ方があるのか，人それぞれで異なるものであるのだろうか？

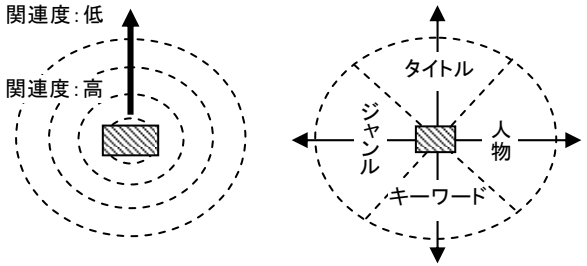
「関連している」というのは，多分に主観的な感じ方である．そこで，我々は，この多分に主観的である「関連している」という事象の定量化を試みた．4 節でこの詳細に関して別途説明する．

3.5 レイアウト算出処理

レイアウト算出処理では，算出された関連度の値が近いほど中央に近く配置されるようにレイアウトを決定する（図 5-(a)参照）．関連度の値をそのまま中央からの距離に割り当てることも可能だが，今回は，リモコンなどでの操作を考え，常に十字キーだけで操作が可能なように，関連度の値が大きいものから順に中央から埋めていく方式とした（配置結果は，図 1 参照）．

中央からの方位方向には，「タイトル」，「人物」，「キーワード」，「ジャンル」のうち，関連度算出処理にて一番優位であった要素（関連度の値の向上に最も起因した要素）で分類し，それぞれ，上，右，下，左方向に配置するようにした（図 5-(b)参照）．

なお，今回は，方位に「タイトル」，「人物」，「キーワード」，「ジャンル」を割り当てたが，これ以外の分類をすることも可能である．例えば，放送局や放送曜日，などで分類することで，新たな一瞥方法となり得る．このように，将来的にはユーザーが方位方向の分類項目を自由に選択出来るようにすると面白い．しかし，今回は，製品化の際の仕様の簡略化，および，リモコンでの操作を考えて，方位による分類は 4 方向（上下左右）とした．この際，後述する外部被験者実験の際に同時に行ったアンケート調査の結果，番組内容の判断材料として優位であった上位 4 項目である「タイトル」，「人物」，「キーワード」，「ジャンル」を選択した．また，特に上位である「人物」，「ジャンル」に関しては，ユーザーが最も操作しやすい方向である横方向に優先的に割り振った．



(a) 距離　(b) 方位

図 5　レイアウト算出処理

4. ユーザーの感じる関連性の定式化

4.1 コンジョイント分析による関連性の定式化

コンジョイント分析 [9] とは，主にマーケティング分野で利用されている実験計画法の一種であり，調査対象者の深層心理を探るデータ解析手法である．商品やサービスの持つ複数の要素について，顧客がどの点に重きを置いているのか，顧客が最も好む要素の組み合わせはどれかを統計的に探るような目的に利用されることが多い．我々はこの手法がユーザーのテレビ番組間の関連性の感じ方という主観量の定量化に応用可能であると考え，コンジョイント分析による関連性の定式化を試みた．

具体的には，被験者に番組情報が記されたカードを 2 枚見せ，この 1 対のテレビ番組の関連性の感じ方を回答してもらう．これを多数のカードの組み合わせに関して行い，そこから番組間の関連性の感じ方に有意な要因を抽出することで定式化を行う．

テレビ番組間の関連性をコンジョイント分析にあてはめるため，テレビ番組情報の各項目要素を属性とし，各項目要素のマッチ具合，つまり類似度のパターンを水準とする．コンジョイント分析では，属性の数は 5 つ前後が適切とされており，むやみに属性を増やすと情報量増加のため回答者が混乱してしまい，かえって非優位な属性にウェイトが集中してしまうことが知られている [9]．また水準も 3 つ前後が適切とされており，多くても 5〜6 水準程度が限界である．このため，番組情報に含まれている全ての項目要素を属性と水準として設定することは出来ない．そこで，調査対象とする番組情報（メタデータ）を簡略化し，定式化に用いる項目要素を限定することで，6 属性 4 水準に収まるように設定した．

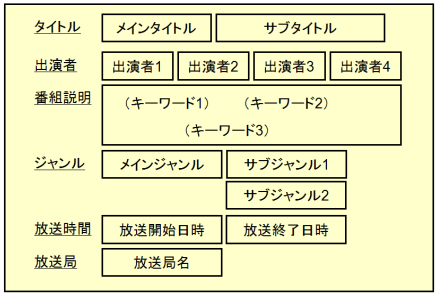


図 6　番組情報の簡略化

番組情報の簡略化の様子を図 6 に示す．番組情報を構成する属性（要素）は，タイトル，出演者，番組内容，ジャンル，放送時間，放送局の 6 つとする．実際の番組情報ではフリーフォーマットで出演者や番組説明が記載されており，出現する出演者数やキーワード数は限定されることはないが，今回の簡略化では出演者は最大 4 名から構成され，番組説明には予め定めたキーワードが最大 3 つまでしか含まれないと仮定する．出演者名，キーワード名は，全ての設問で設定した中から同じものを用いる．つまり，通常の番組情報と異なり，特定の人名，キーワードしか設問中には存在しない．またジャンルは最大 2 つまで付与されず，その 2 つのメインジャンルは同じであ

Vol.2010-HCI-137 No.8
2010/3/19

ると仮定する（実際にはジャンルは 3 つまで付与され，メインジャンルも異なる）．タイトルも，1 組のメインタイトルとサブタイトルとから構成され，設問中では，これらの組み合わせが変化するのみに限定する．

設定した属性と水準は表 1 に示す通りである．表 1 に示された水準を様々に変化させて設問を作成する．これにより，被験者に呈示すべき様々なパターンの番組情報の対が生成される．

表 1　コンジョイント分析の属性と水準の設定内容

属性（要素）	水準（要因）			
	Level 1	Level 2	Level 3	Level 4
タイトル	全一致	メインのみ	1 フレーズ	一致なし
出演者	全員	3 人	1 人	0 人
番組内容	3KW 一致	2KW 一致	1KW 一致	一致なし
ジャンル	両者一致	片方一致	メインのみ	一致なし
放送時間	完全に一致	同じ時間帯	一致なし	
放送局	一致	系列が同じ	一致なし	

本来は，全ての組み合わせに関して顧客に設問を呈示することが望ましいが，そうすると被験者は数千通りの組み合わせに関して回答する必要があり，非常に回答負荷が高く現実的ではない．そこで，コンジョイント分析では，直交配列法という，なるべく少ない設問呈示パターン数で多くの設問を呈示した場合を推定可能とする手法を用いて，呈示すべきパターンを削減する．今回の場合，6 属性（要素）4 水準（要因）の直交配列法によりサンプリングを行い 24 パターンの設問を作成した. 上記のように番組情報を簡略化した場合（表 1）でも，総当たりだと 4096 通り必要な設問を，わずか 24 設問に回答するのみで良くなり，被験者にとって現実的な設問数となる．

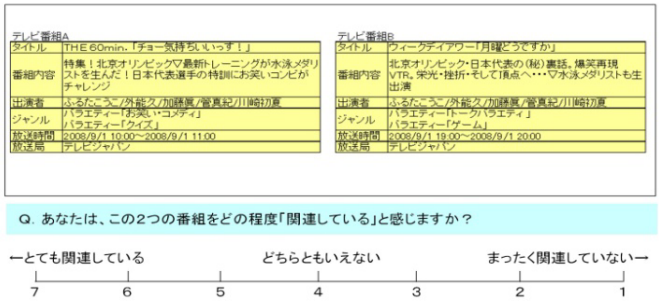


図 7　設問画面の例（簡略化したダミーの番組データ）

被験者には，設問毎にテレビ番組情報が記載された 2 枚のカードを呈示し，被験者に呈示した 2 つの番組情報の関連性に関して 7 段階で評価をしてもらった．カードに記載する番組情報は，ユーザーが予め持っている知識や嗜好等の影響がないように，現実には存在しないダミーのテレビ番組のものとした．設問画面の例を図 7 に示す．

被験者は全国の 20～60 代の男女計 1,800 人であり，年代，性別毎の被験者数は M1（男性 20～34 歳），M2(男性 35～49 歳)，M3(男性 50 歳以上)，F1（女性 20～34 歳），F2（女性 35～49 歳），F3（女性 50 歳以上）の各セグメント 300 人ずつの均等回収である．パソコンを用いたインターネット調査方式（Web アンケート方式）とし，各被験者の実験場所，実験所要時間は問わないものとする．

被験者実験の結果を用いて属性の重要度と水準の効用値を重回帰分析に基づいて算出する．この際，関連度 R の算出式として，式(1)における関数 f を線形式であると仮定する（式(2)）．

$$R = \sum_{n=0}^{M} a_n S_n \tag{2}$$

ここで，S_n は各項目要素における類似度を，a_n はその項目要素における重み要因を示しており，M=6 である（例えば n=1 がタイトル要素，n=2 が人物要素，…，とする）．

このように仮定すると，重回帰分析の結果により得られた属性の重要度は a_n に，水準の効用値は類似度に相当する．以上により，コンジョイント分析を用いることで，人の感じるテレビ番組間の関連性が式(2)として定式化される．

4.2　コンジョイント分析による算出式の評価

関連度算出式を定量的に評価するため，ユーザーカバー率を式(3)で定義する．

$$\text{ユーザーカバー率}(\%) = \frac{\text{納得しているとみなすユーザー数}}{\text{全ユーザー数}} \times 100 \tag{3}$$

納得しているとみなすユーザー数は式(4)により算出する．ここで，Y_i はコンテンツ i に対する被験者回答値の順位，$\hat{Y}_i$ はコンテンツ i に対する関連度算出式が算出した順位，Y_{R_i} は再測定（被験者内の変動要因を吸収するため，同じ被験者に対し同内容の実験を期日をおいて 2 回測定している）したコンテンツ i に対する被験者回答値の順位，$\widehat{Y_{R_i}}$ は再測定したコンテンツ i に対する関連度算出式が算出した順位である．また，ρ_s は閾値であり今回は 0.75 を用いた．

$$\frac{\sum_{i=1}^{N}(\overline{Y_i}-\overline{\hat{Y_i}})(Y_i-\overline{Y_i})}{\sqrt{\sum_{i=1}^{N}(\overline{Y_i}-\overline{\hat{Y_i}})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\overline{Y_i})^2}} \Big/ \frac{\sum_{i=1}^{N}(Y_i-\overline{Y_i})\left(Y_{R_i}-\overline{Y_{R_i}}\right)}{\sqrt{\sum_{i=1}^{N}(Y_i-\overline{Y_i})^2}\sqrt{\sum_{i=1}^{N}(Y_{R_i}-\overline{Y_{R_i}})^2}} \geq \rho_s \tag{4}$$

式(4)は，ユーザー内の主観評価の変動要因を除去した上で，実験により得られたユーザーの評価した関連性の値と関連度算出式により算出された関連度の一致度合いがある一定の値以上であるユーザーを納得しているとみなすこと表している．

Vol.2010-HCI-137 No.8
2010/3/19

コンジョイント分析で定式化した算出式を用いてユーザーカバー率を算出したところ 72.5%となった．つまり，コンジョイント分析を用いた関連度算出式は，上述した条件のもとでは，約 7 割のユーザーの関連性の感じ方をカバーすることが可能であるとみなすことができる．

4.3 コンジョイント分析による算出式の問題点

以上で求めた関連度算出式の実際の番組情報データでの性能評価を行うため，地デジ，BS，CS 放送の実際の番組情報から作成した設問 24 問を用いて，上記と同様の被験者実験を実施した．被験者数は 600 人で同様に各セグメントの均等回収である．図 8 に設問画面の例を示す．簡略化した番組情報と比較して，現実的にはフリーフォーマットで記載された番組内容には多数の人名やキーワードなどが含まれ非常に複雑なデータとなっていることが分かる．

実データを用いた被験者実験の結果，式(3)で算出されるユーザーカバー率は 35.7%と著しく低下することが判明した．これは想定していた値よりも大幅に小さい値であり，算出式の大幅な改善が必要であるといえる．

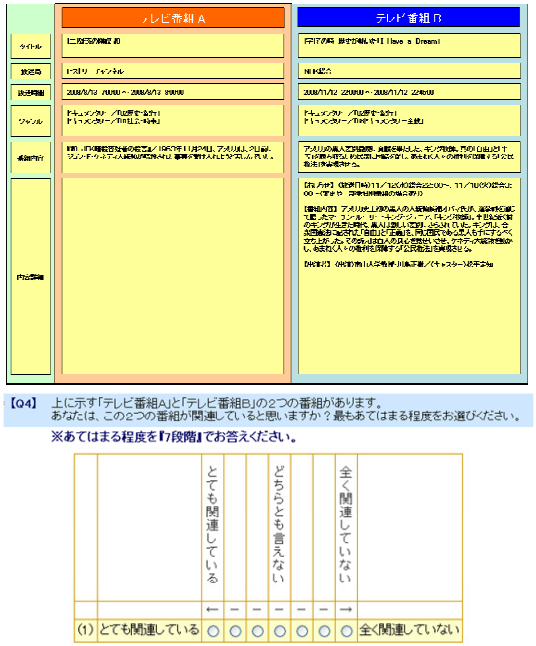


図 8　設問画面の例（実際の番組情報データ）

この要因分析を行ったところ，阻害要因は図 9 の通りであることが分かった．

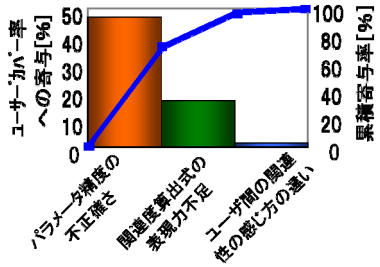


図 9　ユーザーカバー率の阻害要因

一番大きい阻害要因は，算出式のパラメータ精度の不正確さであり，これはコンジョイント分析のために，非常に簡略化した番組情報（4.1 節，図 6）を用いたためである．また，ユーザーの知識や嗜好等の影響がないようにダミーの番組情報を用いたこともこの一因と考えられる．次いで大きな要因は，関連度算出式の表現力不足である．これは，コンジョイント分析の際に，算出式を式(2)のとおり線形であると仮定して直行配列法に基づいて設問を作成したためであると考えられる．先の実験に用いた番組データは属性（項目要素）間の直行性を仮定できたが，実際の番組データはその限りでなく，その要因が想定以上に効いてしまっているといえる．文献 [9] によると，直交配列によって設問を作ると中には非現実的な組合せも出てくるが，パラメータの推定値はかえって現実的になる，との過去の知見も得られているが，テレビ番組間の関連性に応用した場合，この効果よりも上述したような阻害要因による性能劣化の方が大きかったと考えられる．

4.4 関連度算出式の性能改善

実際の番組情報データにて関連度算出式の性能を向上するには，ユーザーカバー率への寄与が非常に大きい上位 2 つの阻害要因を解消すれば良い．この 2 つの阻害要因を解決するために我々は，

① コンジョイント分析に替わるアンケート調査実施手法として，項目応答理論を用いた新たなアンケート調査実施手法の考案と，その手法を用いた実データによる被験者実験の実施
② 関連度算出式の表現能力を向上させるため，実験計画法，遺伝的アルゴリズム，重回帰分析を組み合わせた関連度算出式の最適化手法の確立
③ 延べ40,000 以上の関連度算出結果の目視評価による項目要素抽出精度などの向上の 3 つの施策を講じた．

上述した通り，コンジョイント分析を選んだ一番の理由は，現実的な設問数の被験

Vol.2010-HCI-137 No.8
2010/3/19

者実験で，複雑な要因が絡み合った事象を定式化可能であることであった．しかし，やはり上述した通り，属性・水準の数に関するコンジョイント分析の制約から，テレビ番組の実データを用いた場合，コンジョイント分析では十分な性能が出ないことが判明した．我々は，これを解消するために，コンジョイント分析に替わる新たなアンケート調査実施手法を考案した．これは，項目応答理論（IRT, Item Response Theory）を用いた手法である [11]. IRT は, TOEFL などの能力テストで用いられる手法であり，異なる設問による各受験者の評価を、同一尺度上で比較できるのが特徴である．被験者実験の際の一番の問題点は，より精度の高い算出式を得るためには，番組間の関連性の感じ方に関して非常に大量の回答データが必要であるということであった.「回答する設問を分担して担当させれば、設問が大量であっても一人あたりの負担は小さくなる」と考えることでこの問題を解消することが可能となるが，設問を複数の被験者で分担すると，被験者間の個人差などに起因するばらつき要因などから，異なる被験者の回答を同一尺度上で比較することが出来なかった．そこで我々は，IRT のアンケート調査への適用可能性を検討し，実験により適用可能であることが統計的に証明された．IRT を用いたアンケート調査実施手法を用いることで，30 人のアンケート回答結果のみで，標準的な被験者の 400 人相当の回答結果を推定可能な精度を得られることが分かった．以上により，テレビ番組間の関連性をより詳細に解析するためのデータ取得実験が現実的な被験者数、設問数で可能となった．

この新規手法を用いて，関連度算出式の性能改善のためのデータ取得を目的とした被験者実験を上記と同様の形式で行った. 被験者数は 600 人であり, 設問数は合計 200 問である. 被験者を 10 グループに分け, 各グループに 20 問ずつ割り当てて実施した．IRT を用いたアンケート調査手法の採用により，被験者 1 人あたりの回答負担は増加させず，合計 200 問のデータを取得できたことになる．しかも，コンジョイント分析の際の制約となっていた関連度算出式の線形性の仮定や, 属性数の限定も解消される．阻害要因分析により，2 つめの要因として関連度算出式の表現能力不足が挙げられているが，これも同時に解消することが可能となる．そこで，関連度算出式を線形式から以下の式(5)のような非線形で属性（項目要素）間の相互作用を含む式に拡張する．

$$R = a_1 S_1^{\beta_1} + a_2 S_2^{\beta_2} + \cdots + a_M S_M^{\beta_M} + a_{1,2} S_1^{\beta_1} S_2^{\beta_2} + a_{1,3} S_1^{\beta_1} S_3^{\beta_3} + \cdots + a_{M-1,M} S_{M-1}^{\beta_{M-1}} S_M^{\beta_M} \quad (5)$$

式(5)を用いた場合，関連度算出式として決定しなければならないパラメータ数が爆発し，最適な関連度算出式の算出が困難であるという新たな問題が発生する．そこで我々は，先の IRT を用いたアンケート調査実施手法による被験者実験結果から，この複雑な非線形式を算出する際に，最適な関連度算出式を得るために，実験計画法，遺伝的アルゴリズム, 重回帰分析を組み合わせた関連度算出式の最適化手法を確立した．

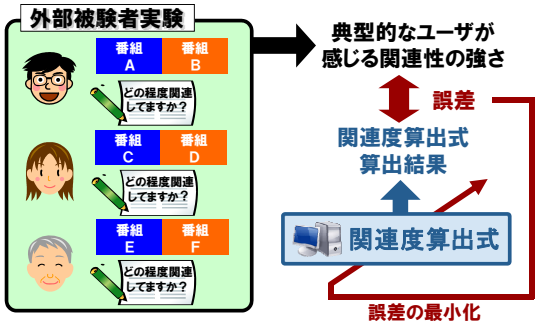


図 10　関連度算出式の最適化

ユーザーカバー率が最も高くなるような関連度算出式を得るのがゴールである．これは，本節で述べた被験者実験の回答結果 200 個を教師データとして用い，4.3 節で説明した被験者実験の結果 24 個を正解データとして用いて, その 2 つの差が最も小さくなるような最適化問題を解くことで得られる（図 10）.

決定すべき項目は，式(5)の a および β である．M=6 であるから，β を 1（単調に増加する）, 2（最初滑らかだがその後急速に増加する）, 1/2（最初急激に増加するがその後定常状態に落ち着く）の 3 通りに限定しても，最適な算出式を求めるための探索範囲は約 150 億通りにもなる．この最適化問題を解くのは非常に困難かつコストが大きく現実的ではない.

そこで，我々は，この探索範囲を絞り込むために，実験計画法，遺伝的アルゴリズム, 重回帰分析を組み合わせて用いた．これにより, 最適化に必要な計算量を約 1/1000 以下に抑えることが可能となった．従来法では数週間必要であった最適な関連度算出式の生成がこれにより 1 日以内に短縮されることになった.

以上によって最適化した関連度算出式のユーザーカバー率を計算したところ，80.8%となった．これは十分に実用的な値である.

実データにおけるコンジョイント分析での関連度算出式のユーザーカバー率は 4.3 節で説明した通り 35.7%と非常に低いものであったが，上述したように大きな阻害要因 2 つを解消するための施策により，80.8%と大きな改善を得ることができた.

なお紙面の都合上，本論文では，①, ②の概要に関してのみ説明した．これらの詳細分に関しては，別の発表の機会に委ねたい.

5. 算出結果の妥当性の評価

5.1 対象番組

地デジ, BS, CS 放送の 1 週間分のテレビ番組（EPG）のメタデータを対象とし，各ジャンル 4 番組ずつ，全 12 ジャンル分で合計 48 番組を注目コンテンツ（検索対象番組）として，関連度算出処理の結果，各軸（タイトル，人物，キーワード，ジャンル）に呈示される上位 10 コンテンツ毎に関して,算出結果の妥当性の評価を行った. 合計，最大 1920 コンテンツが評価対象となる．注目コンテンツ 48 番組は，株式会社ビデオリサーチ社の視聴率調査 [8] を参考とし，同じ番組が重複しないという条件のもとで，各ジャンルにて視聴率の高いものから選択した．

5.2 評価方法

算出結果のコンテンツを 6 名の被験者が分担して評価する．各番組における算出結果が，必ず異なる 2 名の被験者によって評価されるようにする．つまり，評価対象となるコンテンツ数は，延べ最大 3840 コンテンツ分となる．

被験者は，呈示された各々の番組の内容を目視で確認し，

- Good：注目コンテンツと強く関連している
- NG：注目コンテンツと全く関連していない
- OK：上記 Good, NG 以外である

の 3 択のいずれかを選択する．そして，評価結果が NG で無いもの，つまり，全体のうち関連があるコンテンツとして被験者に許容されたコンテンツの割合を適合率と定義する．適合率の全体平均値，および，タイトル軸，人物軸，キーワード軸，ジャンル軸での各々の平均値を評価する．

5.3 評価結果

2009 年 7 月 13 日に取得した地デジ, BS, CS の 1 週間分のテレビ番組のメタデータをオープンデータとして用いて評価を行った．結果，適合率は全体平均で 96.9%，軸毎に見ると，タイトル軸，人物軸，キーワード軸，ジャンル軸でそれぞれ 99.2%, 100%, 88.6%, 99.8%という非常に高い結果となった．

コンジョイント分析で作成した式での結果は，全体で 89.7%，最も低いキーワード軸で 77.1%であったため，関連度算出式の最適化の効果が非常に出ていることが確認できた．

6. 関連研究

井出ら [5] は，関連情報の視覚化手法として極座標を用いて関連する番組を可視化する GUI を提案しており，大坪 [12] は，キーワードや時間による類似度に基づいてビデオコンテンツを連想検索するシステムを提案しているが，いずれも，ユーザーの感じる番組間の関連性という主観量を再現している訳ではないし，可視化表現も異なっている．視聴中番組の番組情報に含まれる出演者名，キーワードを検索キーとして検索する機能を有するデジタル機器はある [10] が，あくまでもキーワードを用いた関連情報検索に留まっており，番組間の関連性を一瞥性高く可視化はしていない．

7. おわりに

本論文では，膨大な映像コンテンツの中から，ユーザーの興味にあったコンテンツを効率良く探すことができるシステムとして我々が開発してきた「ローミングナビ」に関して論じた．ローミングナビは，“関連性”を頼りに関心のあるコンテンツを渡り歩く新しい視聴スタイルを目指したもので，ユーザーの感じるテレビ番組間の関連性という極めて主観的な感覚を，コンジョイント分析，項目応答理論など様々な統計手法を応用して定式化したことが特徴である．これにより，最終的に適合率 96.9%，ユーザーカバー率 80.8%という高い精度でユーザーの感じるテレビ番組間の関連性を再現することができた．ローミングナビは 12 月に発売された当社の液晶テレビ CELL REGZA 55X1 [7] に搭載されている．この新感覚のナビゲーションシステムを是非体験して頂きたい．

参考文献

1) 朝日大学マーケティング研究所, テレビ番組表に関するマーケティングデータ(1), http://reposen.jp/2165/8/66.html, http://www.asahi-bplan.com/marketing/data/0310.pdf, (2003).
2) 市村他, 質問応答と日本語固有表現抽出および固有表現体系の関係についての考察, 情報処理学会研究報告, NL-161-3, (2004).
3) 山崎他, 話題抽出エージェントを用いた番組検索システムの実装, コンピュータソフトウェア, Vol. 25, No. 4, pp. 41-51, (2008).
4) コンテンツ指向ユーザーインタフェース“ 関連ナビ ”,東芝レビュー, Vol.64, No.3, p.4, (2009).
5) 井出他, 関連情報の視覚化手法の提案 – 線表現による関連性明示と表示効率化の両立 –, 信学技報, vol. 109, no. 75, MVE2009-17, pp. 91-94, (2009).
6) CEATEC JAPAN 2008 東芝プレスリリース, http://www.toshiba.co.jp/about/press/2008_09/pr_j2601.htm, (2008).
7) CELL REGZA ローミングナビ機能紹介ページ, http://www.toshiba.co.jp/regza/lineup/55x1/recording.html, (2009).
8) 株式会社ビデオリサーチ, 視聴率調査, http:// www.videor.co.jp/data/ratedata/r_index.htm
9) 朝野熙彦，入門 多変量解析の実際，講談社，ISBN4061539469，(1996).
10) ソニーレコーダ機能紹介ページ，http://www. sony.jp/bd/lineup/popup/05.html
11) 高橋正視，項目反応理論入門 － 新しい絶対評価，イデア出版局，ISBN 4900561002，(2002)
12) 大坪五郎，Goromi-TV 撮りためた千以上のビデオを気ままに閲覧する方法, WISS2006, pp. 47-52, (2006).