番組広報によるＴＶ番組のクラスタリング

浦谷 則好　　山田 一郎
NHK 放送技術研究所
{uratani,ichiro}@strl.nhk.or.jp

1．はじめに

デジタル放送時代に入り、マルチメディア化や多チャンネル化が進んでいる。放送番組にもＥＰＧ（電子番組案内）に限らず、文字情報による番組内容が付加されていれば視聴者にとって便利である。また、映像データベースを利用する際にも検索のキーとなるのは内容記述文やそれから得られるキーワードになることが予想される。しかし、どんな情報が検索者にとって有効なのか、またどれくらいの検索精度が得られるものなのかはいまだに明らかとなっていない。

番組検索のためのキーワードとして何が有効なのか、どの程度の分類が可能なのかを把握するため番組広報を用いてクラスタリングを実施したので、それについて報告する。

2．キーワードのクラスタリング

ＮＨＫが新聞社や出版社に番組宣伝のために送っている資料が番組広報である。（図１）286番組分の番組広報を収集し、番組内容記述部を形態素解析し、キーワード（自立語）を自動抽出した。収集した番組の一覧を表１に示す。表以外では２回分のものが９番組、１回分のものが34番組である。抽出されたキーワードは6527個（異なり数）であった。１回しか出現しなかった4355個のキーワードと、出現頻度は多いが分類の役に立つと思われない「番組」と「ほか」の２語を除き、残りの2170個をすべてキーワードとして選定した。１回しか出現しないキーワードはその番組を他と区別するためには有用だが、他の番組との距離を測るには役に立たないからである。これらのキーワードをそのまま使って番組間の距離（あるいは類似度）

表１．　クラスタリング対象の番組

ベストオブクラシック	２１
ＥＴＶ特集	１８
ひるの歌謡曲	１５
ＮＨＫスペシャル	6
Ｎ響アワー	5
すこやかシルバー介護	5
ためしてガッテン	5
ふれあい通り	5
コメディーお江戸でござる	5
トップランナー	5
共に生きる明日	5
金曜時代劇	5
週刊ボランティア	5
食卓の王様	5
水曜シリーズドラマ『鏡は	5
青春探検	5
土曜ドラマ	5
未来潮流	5
夢用絵の具	5
ときめき夢サウンド	4
なぞ解き歳時記	4
ふたりのビッグショー	4
ふるさとの仲間たち	4
アジア発見	4
セッション97	4
バラエティー生活笑百科	4
海外ドキュメンタリー	4
小さな旅	4
生きもの地球紀行	4
大河ドラマ『毛利元就』	4
堂々日本史	4
ＢＢＣロックライブ	3
ＦＭシアター	3
ＮＨＫ歌謡コンサート	3
ことばてれび	3
ふるさとの伝承	3
クイズ日本人の質問	3
ポップスグラフィティ	3
芸能花舞台	3
世界・わが心の旅	3
世界の民族音楽	3
知への旅	3
地球ロマン	3
中学生日記	3
土曜元気市	3
土曜特集	3
俳句王国	3
未来派宣言	3

を計算することも考えられるが、以下の２つの理由で好ましくないと考えられる。

（１）分類の基準となる次元（この場合にはキーワード）は独立であることが望ましいが、キーワード間には共起関係があり、これを満たさない。

（２）クラスタリングに要する計算時間を考慮すれば次元の数は少ないほど好ましい。

その上、キーワードのクラスタリングを行えば、既存のシソーラスでは得られない対象分野依存のキーワードの関係を得ることができる。そこで、まずキーワードのクラスタリングを実施した。

各番組をベクトルの次元にとり、次式でキーワード間の類似度を定義した。

$$sim(w_i, w_j) = \frac{2 \times |w_i \cap w_j|}{|w_i| + |w_j|}$$

ここで$|w|$はキーワードwが番組（広報）中に出現する回数、$|w_i \cap w_j|$はw_iとw_jが共に出現する回数である。ただし、番組広報の文章は短いので１番組にキーワードwが複数回出現しても、１回とカウントした。2170個のキーワードを286次元で分類する程度であればそのまま実施してもそれほど演算時間は掛からないが、将来もっと大規模な実験を実施することを考え、Brown,P.F.らが採った方法[1)]のように初期クラスターを適当に設定し、１つずつデータを追加する毎に、最も類似したデータをマージすることでクラスター数を維持しながらクラスタリングする逐次的な手法を用いた。全てのデータ入力が終了した後のクラスタリングは通常の階層的クラスタリング手法と同じである。初期クラスターの数とその選定法を変えて、最終的に作られたクラスターを「初期クラスター数2000を頻度の降順に選んだもの」との差異（適合率、再現率）を比較したのが表２である。類似度が0.5を超える場合のクラスタリングでは、（表２に挙げなかったものも含めて）初期クラスターが1700以上ではクラスター数や選定法によってほとんど差異が生じないことが分かった。しかし、クラスターにまとめる類似度を0.5以下に設定した場合には初期クラスターの影響はかなり大きい。当然のことながら初期クラスター数は大きいほど差異が小さくなる傾向がある。降順・昇順の別では昇順に初期クラスターを選ぶ方が影響が小さい。これは、クラスターが統合されるとき、頻度の小さいものが統合されても他のクラスターとの類似度はあまり変化しないのに対し、頻度の大きなものでは影響が大きくなるためと想像される。

以上の結果から、クラスタリング処理を類似度0.5を超える範囲で止めることにした。得られたクラスターの数は1538個であった。最大のクラスターは42個のキーワードから構成されていたが、1289個のクラスターは１つのキーワードしか含んでいなかった。キーワードを多く含むクラスターを示したものが表３である。含まれるキーワードを見ると番組名を推定することができ、分野依存のキーワードの共起関係を予想どおり抽出することができていることが分かる。Ａは「食卓の王様」、Ｂは「毛利元就」、Ｃは「ためしてガッテン」、Ｄは「コメディーお江戸でござる」、Ｅは「ＢＳ日曜スペシャル『ロシア共産党は・・・』」であり、

表２．　初期クラスターの違いによる結果の差異

初期クラスター数・選定法	類似度0.501		類似度0.333	
	prec	recall	prec	recall
２０００・Ｒ	.996	.997	.846	.886
１７００・Ｓ	.995	.996	.840	.865
１７００・Ｒ	.996	.996	.853	.875
１５００・Ｓ	-------	------	.850	.885
１５００・Ｒ	-------	------	.854	.877
１０００・Ｓ	-------	------	.841	.875
１０００・Ｒ	-------	------	.814	.853
７００・Ｓ	-------	------	.703	.809
７００・Ｒ	-------	------	.850	.885
５００・Ｓ	-------	------	.703	.759
５００・Ｒ	-------	------	.749	.826

Ｓ：頻度の降順　Ｒ：頻度の昇順
最終クラスター数　1538個/0.501　474個/0.333

これ以外にも「クイズ日本人の質問」、「ためしてガッテン」、「生活笑百科」、「水曜シリーズドラマ『鏡は眠らない』」、「金曜時代劇」、「俳句天国」など出演者が固定している番組でキーワードがクラスタリングされる傾向が見られた。

表3．　クラスターの例（大きい順に5個）

A (42)	食卓、フライパン、和菓子、カレー、オング、王様、国井、ジュディ、・・・
B (31)	渡辺、月山、晴久、恵、大内、平野、上川、毛利、永井、銀山、城、・・・・
C (27)	紀行、腰痛、脂肪、常識、大仁田、筋肉、田崎、松本、ツボ、ワイン、・・
D (27)	江戸、由紀、細川、コメディー、講座、四朗、風俗、徳治、三太郎、桜、・・・
E (23)	改革、崩壊、政界、エリツィン、混迷、セマゴ、大統領、共産、巨万、・・・・

3．番組のクラスタリング

2．で抽出したクラスターをもとに、番組のクラスタリングを実施した。すなわち、同一のクラスターに属するキーワードは同一と見なし、各番組を1538次元のベクトル（キーワード）で表現し、類似度が高いものを統合し、代表ベクトルを更新するというやり方で、286番組が1つのクラスターになるまで階層的クラスタリングを行った。類似度の定義式は番組とキーワードの役割が逆転するだけで2．で用いたものと同じである。

得られた樹形図を図2に示す。類似度が0.5以上で統合できたのはわずかに3回（「食卓の王様：ゴマ」ー「食卓の王様：開け開けゴマパワー」(0.810)、「ＢＳ日曜スペシャル：ロシア共産党は混迷を救えるか」2回分(0.769)、「小さな旅：青春の車窓」ー「小さな旅：忘れられない人」(0.507)）であった。反対に類似度0.01以下で無理やり統合したのは、図2の上位の2つ「ベストオブクラシック：リヨン・・・」(0.007)と「スクール五輪の書」(0.003)である。こうしたクラスタリングの結果を客観的な評価尺度がないので評価が難しいが、主観評価では田中の手法[2]（級間変動が最大になるところによって分割を繰り返す）に比べて、かなり良好な結果が得られている。木の深さは最大で57で、田中の方法（最大149）と比べてバランスの良い木が作られていることが分かる。図2からも分かるようにドラマ、音楽、芸能などがきれいにクラスターを形成している。反対に、同種の番組でありながらクラスターを形成しなかったものとしては、「青春探検」、「トップランナー」、「共に生きる明日」などが挙げられる。これらは出演者やテーマが毎回変わるため、キーワードだけでは似ていると判断できないからと思われる。また、金曜時代劇の2回分が中学生日記と同じクラスターに統合されているのが興味深い。この回は寺子屋に通う子供の問題を取り上げていて、内容はまさに時代劇版の「中学生日記」となっている。

4．おわりに

番組広報を元にキーワードのクラスタリングと番組のクラスタリングを実施し、有用な結果を得ることができた。しかし、データが大きくなった場合は、計算時間が田中の方法に比して莫大となるため、計算時間の短縮法を検討することが必要である。また、今回は個々のキーワードの頻度を考慮しなかったが、考慮した場合はどうなのか、類似度のかわりに内積を用いたらどうなのか、などを比較・検討していきたいと考えている。また、昨年報告した判別法[3]との融合や使い分けについても検討を加えていく予定である。

参考文献

1）Brown,P.F. et al.:Class-Based n-gram Models of Natural Language, Computer Linguistics,Vol.18,No.4(1992)
2）田中英輝：大規模文書集合の高速クラスタリング，言語処理学会第3回年次大会pp.249-252，(1997)
3）浦谷則好ほか：ＮＨＫニュース記事からのキーワード抽出と記事の分野判別実験，言処第3回年大，(1997)

七月十九日（土）総合・後11・00～11・25

☆アジア発見

「大草原の父と街の息子」

～中国・内モンゴル自治区～

中国・内モンゴル自治区は、人口の13％を少数民族が占めている。広大な草原では何千もの昔から遊牧民が羊を追い、移動を繰り返してきた。

サインバトルさん（38）もこの草原に生まれ、遊牧生活をしてきた。しかし、13年前にテントでの移動生活をやめ、煉瓦の家で定住を始めた。政府の定住化政策によって土地が小さく区分され、移動範囲が狭められたことがきっかけであった。

定住がもたらしたものは、頑丈な家と便利な暮らし。さらには、都市との距離も縮まった。サインバトルさんは今、教育のために子どもを都市に住まわせ、親子離れ離れの生活を送っている。

その一方で、過剰放牧による草原の砂漠化も進んでいる。柵を立てたり種をまいたりして生産性を上げているものの生計が立たず、使用人になる者も出てきている。

定住化が草原の家族にもたらしたものとは何なのか。内モンゴルで消え行く遊牧民の生活を伝える。

※当初7／12(土)放送予定が「神戸・小学生殺害事件容疑者逮捕」関連ニュースなどにより延期。

担当（大阪局・文化）大井
（大阪局・広報）土橋

図1．　番組広報の例

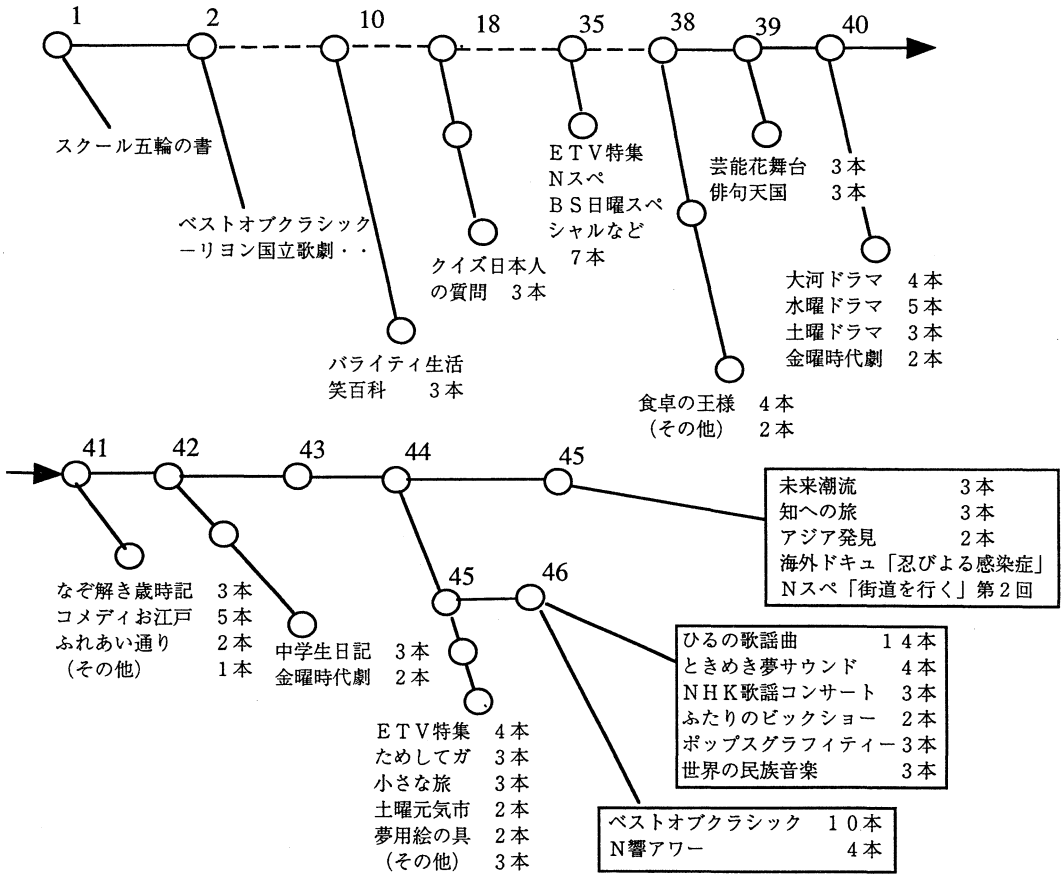

図2．　番組クラスタリングの結果（一部）