ニュース原稿からの話題抽出を利用したテレビ番組選択の検討

2P－7　山田一郎　金淵培　柴田正啓
NHK放送技術研究所

1. はじめに

デジタル放送では、ニュースや一般番組に、その内容を説明したインデックスを付加して家庭まで送るサービスが検討されている。このインデックスと視聴者の個人情報を利用すれば、視聴者の嗜好に合った番組を視聴者側で自動選択することができる。しかし、従来から提案されている個人情報を利用した番組選択手法[1]では、選択される番組は個人情報の記述内容だけに依存するため、いつも同じような番組が選択され、新たな種類の番組選択が難しい。また、最近どのような話題があるか、何が人気となっているか、といった一般的な項目も考慮されず、放送サービスとしては十分でない。

そこで本稿では、世間で話題となっている出来事を抽出し、利用することによって、自動番組選択の幅を拡張する手法を提案する。このために、世間の動きを反映した情報を含むニュースを利用して、特定期間における主要な話題を抽出し、その話題に関連する番組を検索する。この処理により、自動番組選択時に、世間で話題となっている出来事に関連する番組の選択が可能となり、番組選択を効果的に拡張することができる。

2. 話題抽出

我々は、アナウンサーに実際に読まれるNHKのニュース原稿を対象にして、1ヶ月ごとの話題を抽出する研究を進めている[2]。このニュースの話題を、放送波またはインターネットなどを介して視聴者へ送ることにより、視聴者側での世間で話題となっている出来事に関連する番組選択が可能となる。

この手法では、まずニュース原稿の第一文に含まれる単語の話題性を評価する。対象月の単語 t の出現頻度を $n(t)$、その期待値を $e(t)$、対象月のニュース原稿の総数を N、その中で単語 t が出現する原稿の数を $df(t)$としたとき、月ごとに変化する単語 t の話題値 $weight(t)$ は以下の式とした。

$$weight(t) = \frac{(n(t) - e(t))^2}{e(t)} \times log(\frac{N}{df(t)})$$

単語の頻度が期待値より小さいときも大きい場合と同じ正の値をとってしまうため、その場合は、$weight(t)=0$ としている。

次に各ニュース原稿を、含まれる単語をベクトルの要素に、その単語の話題値を要素の値とした特徴ベクトルで表現する。このとき、2つの特徴ベクトル間の類似度($0 \leq$類似度≤ 1)を以下の式で定義する。

$$\text{類似度} = \frac{\text{共通する要素の値の和}}{\text{2つの特徴ベクトルの要素の値の和} - \text{共通する要素の値の和}}$$

各原稿を単一の要素からなる初期クラスタとし、類似度が最大のクラスタを同一クラスタに結合する作業を、最大の類似度が一定のしきい値（0.33)以下になるまで繰り返すことにより、似た話題を持つ集合にクラスタリングする。クラスタを特徴づける重心ベクトルは、クラスタに含まれるニュース原稿中の単語をベクトルの要素に、その単語の話題値と単語のクラスタ内での出現率の積を要素の値とした。

最後に各クラスタから、代表する名詞句を抽出し、重心ベクトルの要素の値の和とクラスタに含まれる原稿数の積の降順に重要と判断し、この月の話題として抽出した。1998年9月の話題抽出結果の上位8項目を図1に示す。

[1]:台風七号の接近(703個)
[2]:金融再生関連法案の修正協議(241個)
[3]:カーディナルスのマーク・マグワイア選手(160個)
[4]:防衛庁の装備品の調達(131個)
[5]:北朝鮮・朝鮮民主主義人民共和国のミサイル発射(186個)
[6]:缶入りのウーロン茶(55個)
[7]:長銀・日本長期信用銀行(59個)
[8]:金融危機の連鎖(149個)
※括弧内の数字はクラスタに含まれる原稿数

図1. 話題抽出実験結果（1998年9月）

A Study on Using Topic Detection from News Corpus for TV Program Selection
Ichiro Yamada, Yeun-Bae Kim, Masahiro Shibata
Japan Broadcasting Corporation
ichiro@strl.nhk.or.jp

3. 話題に関連する番組検索

ニュースで大きな話題となった項目は、一つの特集番組として放送されることが多い。実際に朝日年鑑(1998年)に記載された話題を対象として調査した結果、45項目の話題中、38項目(84%)が、NHKの番組で取り上げられていた。つまり、話題に関連している番組は多く存在する。

しかし従来のキーワード検索では、的確に番組を検索することは難しい。そこで本手法では、検索条件となるニュースの話題として、前章で述べた話題を形成するクラスタの重心ベクトルを利用する。図2に話題のベクトルの例を示す。

検索対象となる番組情報は、NHKの番組広報用に生成された番組説明文を利用する。番組説明文の例を図3に示す。この説明文を形態素解析し、そこに含まれる単語を利用して、話題との類似度(0≦類似度≦1)を以下の式で定義する。

$$\text{類似度} = \frac{\text{番組情報と共通の話題ベクトルの要素の値の和}}{\text{話題ベクトルの要素の値の和}}$$

613個の番組情報を対象として、話題からの検索を行った。話題と全ての番組との類似度を求め、類似度が一定のしきい値(0.5)以上の番組を検索結果とした。図4に1998年9月の話題の上位8項目から検索した結果を示す。人手により検証を行ったところ、検索対象の番組情報には、この月の話題の上位8項目に関連した番組は他に無く、良好な検索結果が得られている。

話題抽出結果の上位の話題に関連した番組ほど、世間の話題を取り上げた番組と判断でき、番組選択時の候補として利用できる。

北朝鮮・朝鮮民主主義人民共和国のミサイル発射

{ミサイル(6879.8), 発射(2932.5), 北朝鮮(2868.2), 朝鮮民主主義人民共和国(1520.6),防衛庁(1313.6), 協議(292.1), 便(162.6), ニューヨーク(114.9), 額賀(110.1), 修正(107.4), 高村(83.5), 偵察衛星(74.7), 開発(66.6), 防衛(66.6), ・・・・・}

※各単語はベクトルの要素
括弧内の数字はベクトルの要素の値

図2. 話題ベクトルの例

NHKスペシャル 1998年08月09日 総合

「核・連鎖の時代へ」インド、パキスタンとなぜあいついで核実験を行ったのか。核拡散を防ぐ新たな手だてはあるのか。核廃絶を訴える被爆国・日本の役割を検証する。

図3. 番組説明文の例

4. まとめ

本報告では、番組選択時の検索条件の拡張のために、月単位で抽出した話題から一般番組を効果的に検索する手法を述べた。この処理により、ニュースと一般番組といった異なるデータベース間のリンクを生成できる。さらに、視聴者がニュースを見ている時に関連番組を宣伝することができるなど、効果的な情報活用にも応用できる。

【参考文献】

[1]矢川ほか「個人の嗜好に合ったテレビ番組を自動編成するエージェントの検討」信学技報,AI98-55,pp.9-16(1998)

[2]山田ほか「ニュース記事を利用したトピック抽出の検討」言語処理学会第5回年次大会論文集,pp116-119(1999)

[3]:カーディナルスのマーク・マグワイア選手
→「クローズアップ現代」 大リーグ史上最強の対決〜マグワイアVSソーサ〜 (9/28) 類似度 0.60
マグワイア選手とソーサ選手。今夜は二人の大リーガーが繰り広げるホームラン競争を伝える。

[5]:北朝鮮・朝鮮民主主義人民共和国のミサイル発射
→「クローズアップ現代」 “テポドン”の衝撃〜検証・北朝鮮ミサイル発射〜(9/7) 類似度 0.71
日本上空を飛び越えて、大平洋に着弾したテポドン。今夜は、北朝鮮が発射したミサイルの衝撃を伝える。

→「日曜討論」 北朝鮮ミサイル発射日本の対応を問う(9/6) 類似度 0.71
北朝鮮のミサイル発射問題について、日本の対応を問う。

[6]:缶入りのウーロン茶
→「特報首都圏」 ’98続発毒物混入事件〜長野・新潟からの報告〜 (9/13) 類似度 0.87
私たちがふだん口にする飲料物に、毒物などが混入される事件が相次いでいる。長野県須坂市のスーパーマーケットに、・・・。長野県須坂市、新潟市で起きた事件を検証し、続発する事件の背景を追う。

図4. トピックに関連する番組の抽出結果（類似度のしきい値0.5）