5C-4

単語間の意味的関係を用いたテレビ番組マップ生成

三浦 菊佳† 山田 一郎† 宮﨑 太郎† 加藤 直人† 田中 英輝†

ＮＨＫ 放送技術研究所†

1 はじめに

放送局では、多種多様な番組が日々制作されており、これらを効果的に管理し、利用することが求められている。現在は、放送日時やジャンルなど表層的な情報による管理が主流で、コンテンツの内容を軸とした整理は行われていない。このため、たとえば「高血圧」に関する番組には、治療法を説明する健康番組のほか、予防法として減塩レシピを紹介する料理番組、対処法として適度な運動を紹介する情報番組など、ジャンルを超えた様々な番組があるが、これらを関連付けることができない。

また、現在、NHK オンデマンド[1]や NHK アーカイブス[2]では、おすすめ番組として、ユーザが選択した番組に関連する番組を提示している。この処理では、EPG(電子番組表)に含まれる番組概要文中の単語の完全一致を手がかりとして類似番組を提示する[3]ため、上記の例のように内容自体が類似していない番組を結びつけることはできない。

以上のような問題点に、より柔軟に対応できる番組管理を実現するために、我々は、任意の番組の関連性を一覧できるテレビ番組マップ(図 1)を提案する。単語の意味的なつながりを介して番組を関連付けられるため、番組管理の拡張のみならず、視聴者向けのコンテンツ推薦サービスにも役立てることができる。本稿では、番組の主題と番組内容との関係を推定し、番組間をリンクで結ぶ手法について述べる。NHK の番組「きょうの健康」を対象として行った実験結果、提案手法の有効性と問題点の検証を報告する。

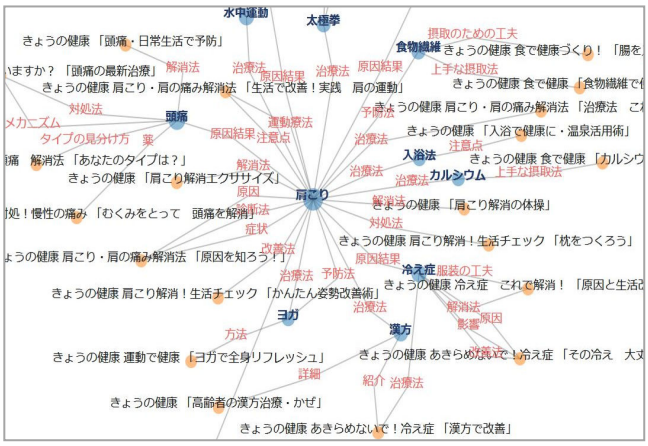

図 1. テレビ番組マップ例 （青い丸が主題を表す単語、オレンジの丸が番組を示す）

TV program map generation using the semantic relations between words

Kikuka MIURA† Ichiro YAMADA† Taro MIYAZAKI† Naoto KATO† Hideki TANAKA†

† Science and Technology Research Laboratories, Japan Broadcasting Corporation

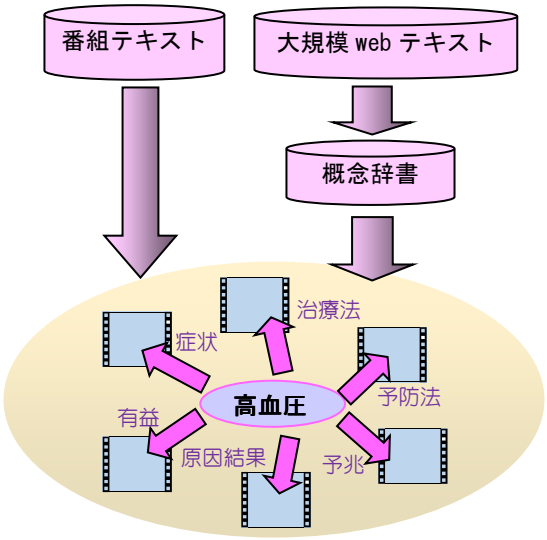

図 2. 番組間リンク付与概要

2 提案手法

提案手法では、番組のサブタイトルと概要文（これらを以後、番組テキストと呼ぶ）、さらに、大規模 Web テキストから抽出した単語と単語の関係を記した概念辞書を入力とする(図 2)。提案手法の処理手順を図 3 に示す。まず、主題を表す単語を番組テキストから抽出する(図 3, 処理 1)。この処理では、サブタイトルと概要文に共通して出現する名詞と、tfidf 値[4]の高い名詞を優先し、主題を表す単語を 1 つ抽出する。

次に、概念辞書を用いて、主題と番組の関係を推定する。ここで扱う概念辞書とは、大規模 web テキストデータから、2 つの単語とその間の関係名を Stijn らの手法[5]により抽出したものである。たとえば、「高血圧を防ぐには減塩が有効」といったテキストから、「高血圧」、「減塩」という 2 単語と、関係名「予防法」を自動獲得し、3 つ組<高血圧, 予防法, 減塩>（以下、関係項目と呼ぶ）の形式で蓄積する。次章で説明する実験では、以下の 10 種の関係名に対して関係項目 249,119 個を抽出し概念辞書として利用した。

［関係名リスト］治療法、症状、予防法、予兆、有益、原因結果、材料、製品、部分、名物

主題を表す単語が含まれる関係項目のうち、関係名が番組テキストに出現する場合(図 3, 処理 2-1)、もしくは、関係項目の他方の単語が番組テキストに含まる場合(図 3, 処理 2-2)、その関係項目の関係名を主題と番組の関係候補とする。たとえば、主題が「高血圧」と推定された番組で「減塩」という単語が番組テキストに含まれていれば、関係項目<高血圧, 予防法, 減塩>から、主題

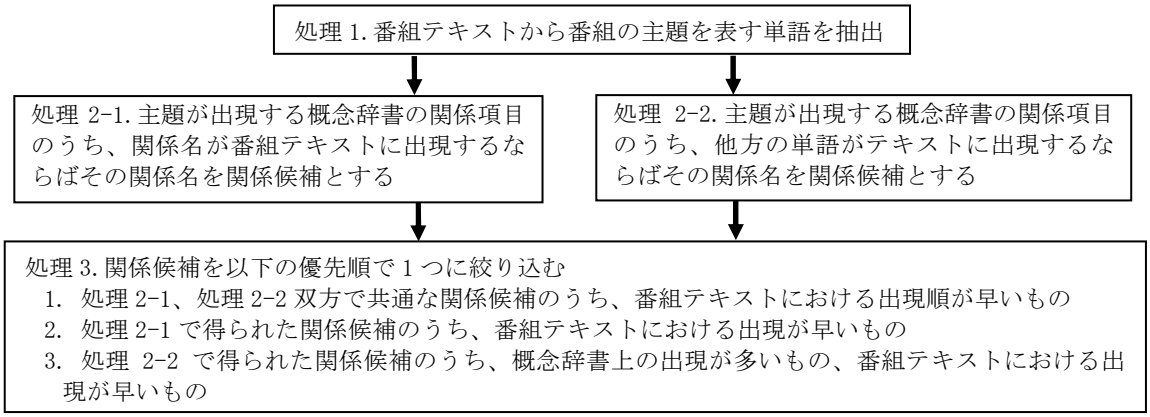

図 3.処理手順

「高血圧」と番組との関係候補として「予防法」を獲得する。複数の関係候補が処理 2-1 と処理 2-2 で得られた場合は、優先順(図 3,処理 3)に従い 1 つに絞る。

3　実験

3－1　実験方法

提案手法の有効性を検証するため、NHK で放送された「きょうの健康」1,053 番組の番組テキストを対象として、主題抽出と、主題と番組間の関係名を推定する 2 つの実験を行った。評価用のデータとして、著者を含まない 3 人のアノテータが正解データを作成した。主題は番組テキストに含まれる名詞から抜き出し、関係名は概念マップに出現する関係名リスト(「その他」を含めて 11 種）から選択した。

1,053 番組中、2 人以上で主題および関係名が一致し、関係名が「その他」でないものは 765 番組であった。これを実験対象とした。

3－2　実験結果

主題抽出の評価結果を表 1 に示す。

表 1. 実験結果（主題抽出）

適合率
72.0% (551/765)

主題抽出処理で誤抽出した場合、関係名推定処理結果にも影響する可能性がある。ここでは関係名推定のみの評価を正当に行うために、主題の正解データを利用して関係名を推定した。評価結果を表 2 に示す。

表 2.　実験結果（関係名推定）

適合率	再現率
60.8% (292/480)	38.2% (292/765)

3－3　考察

主題抽出では、正解を一つに限定する厳しい条件にも関わらず、適合率が 72.0%と手法の有効性を確認できた。実際は、不正解と判定されたものでも主題として相応しいものも存在しており、番組管理やコンテンツ推薦システムなどで利用する処理としては十分な結果が得られている。

関係名推定では、提案手法で 1 つも関係名を獲得できない番組が 285 番組存在した。そのうち主題の単語自体が概念辞書に存在しないものが 85 番組、主題の関係項目は概念辞書に存在するが関係名を獲得できなかったものが 200 番組あった。概念辞書に存在しても獲得できないケースでは、同義語辞書を参照することで、抽出数を上げられると考えられる。たとえば、概要文に「脳卒中のサイン」という記述がある場合、「サイン」が「予兆」と同義であると特定できれば「予兆」という関係名が獲得できる。関係名推定処理で誤推定した番組を精査したところ、69 番組は複数の関係候補から絞り込む過程で誤判定していた。今回はアドホックな優先順位により候補の絞り込みを行ったが、今後、優先度の条件を再検討する必要があると考えられる。

4　まとめ

本稿では、テレビ番組マップ生成のための、番組の主題と番組との関係を推定する手法を提案した。「きょうの健康」765 番組のサブタイトルと概要文を対象に実験を行った結果、一定の効果を確認した。今後、適合率、再現率を上げるとともに、番組推薦システムなどに応用していく予定である。

参考文献

[1]NHK オンデマンド https://www.nhk-ondemand.jp/
[2]NHK アーカイブス http://www.nhk.or.jp/archives/
[3]Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu and Mike Gatford. Okapi at TREC-3. (TREC 1994).
[4]徳永健伸,情報検索と言語処理,1999.
[5]Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda and Masaki Murata. Large Scale Relation Acquisition using Class Dependent Patterns. (ICDM'09).