D-006

視聴中の番組を起点とした関連番組検索

TV Program Retrieval based on Information of the Program Being Viewed

後藤 淳†‡　住吉英樹†　宮崎 勝†　柴田正啓†　相澤彰子‡
Jun Goto　Hideki Sumiyoshi　Masaru Miyazaki　Masahiro Shibata　Akiko Aizawa

1. はじめに

近年、NHK オンデマンドなど、インターネット等を通じて番組を提供するサービスが普及し始めている。我々は、このような番組データベースから自動で関連する番組を検索する TV システム及び Web サービスを検討している[1]。本稿では、視聴中の番組の内容を説明したテキスト情報（EPG、字幕等）を利用して、番組内容が関連する番組を検索する手法を提案したので報告する。

2. 関連研究

コンテンツの検索・提示に関連するサービスとして、You-tube やニコニコ動画などの動画サイトがある。これらのサービスでは、ユーザの投稿した動画が主に対象となるため、内容を説明した文章がコンテンツに付与されておらず、ユーザ付与のタグや選択履歴に基づき関連動画の提示が行われている。また溝口らは、6 種類の簡易オントロジに基づく類似度を定義し、番組の関係をグラフとして提示する TV システムを提案している[2]。しかし、使用するオントロジは、放送局や放送時間、ジャンル、出演者など特定の語彙を対象としており、番組内容自体の類似性については考慮していない。また、山崎らは、EPG の情報に対する複合語処理と 4000 語の意味辞書により話題ラベルを番組に付与し、ラベルに基づく関連番組検索を提案している[3]。しかし、ラベル候補はルールや辞書に基づき抽出する固有表現を含むものであり、EPG の文章にラベル候補がない場合も考えられる。

本研究では、番組の内容を説明している EPG の番組概要に加え、番組紹介 Web ページやクローズドキャプションを利用し、番組の内容に基づく番組検索を行う。事前に特定のラベル付与は行わず、固有表現と単語 n-gram により重み付けした Okapi BM25 尺度により関連番組を検索する。

3. 番組内容を説明するテキスト情報

本研究では、番組の内容を説明するテキスト（番組関連テキスト）として、以下の情報を利用する。

- **電子番組表 (EPG)**

デジタル放送では、タイトルや放送時間、出演者のほか、番組の内容を説明した概要が付与されている。ただし、付与される番組紹介文の質や量は、放送局や番組ジャンルにより異なる。ドラマやドキュメンタリには、あらすじ等を記述した詳細概要が付与されている。

- **番組ホームページ (HP)**

各放送局では、放送する番組や放送済みの番組の内容に関する情報を放送局のホームページ(HP)で紹介している。例えば、クローズアップ現代など時事の話題を放送する番組では、放送時に EPG に付与する情報は少量であるが、放送後に HP に詳細が掲載されている。そのため HP の番組紹介を EPG の補完情報として用いる。

- **クローズドキャプション (CC)**

ニュースや報道番組は、リアルタイム性が要求されるため、必ずしも詳細な内容が EPG に付与されない。一方、定時ニュースなど、アナウンサの声を文字化したクローズドキャプション(CC)が付与される番組が増加している。そこで、ニュース番組等の内容を把握するテキスト情報に CC を用いる。CC を検索元情報に利用することで、ニュースの項目毎に、関連番組を検索することができる。

4. 提案手法

本研究では、番組間の関連度を、3. で述べた番組関連テキストの類似度として定義する。タイトル、ジャンル、出演者などのメタ情報に基づく番組の関連性は対象としない。

4.1 関連番組検索

検索元コンテンツ Q があったときの番組データベース DB 内のコンテンツ D との関連度スコア S(D,Q)を(1)式で定義する。T_n は Q に含まれている単語 n-gram(n=1, 2, 3)とする。対象とする n-gram は予め定めたストップワード（助詞等）を含まないものとする。

$$S(D,Q)=\sum_{n}\sum_{T_n\in Q} w_{ene}(T_n)\times w_{ng}(T_n,n)\times S_{BM25}(T_n,D) \quad (1)$$

S_{BM25}、W_{ene}、W_{ng}について、以下に説明する。

- **Okapi BM25 に基づくスコア (S_{BM25})**

Okapi BM25 [4]に基づき、対象番組 D における T_n のスコア $S_{BM25}(T_n, D)$ を (2)式で定義する。なお、パラメータ K は(3)式のとおりである。BM25 は、パラメータ調整により対象のデータに適応できる。今回、k_1=3.0, b=0.75, k_3=100 と設定した。

$$S_{BM25}(T_n,D)=\frac{(k_1+tf)}{K+tf}\frac{(k_3+1)qtf}{k_3+qtf}\log\left(\frac{M-m+0.5}{m+0.5}\right) \quad (2)$$

$$K=k_1((1-b)+b\times dl/avdl) \quad (3)$$

M : DB の番組総数、m : T_n を含む番組数、tf : D 中の T_n の頻度、qtf : Q 中の T_n の頻度、dl : D の説明文長、avdl : DB 内の全説明文の平均長とする。

†NHK 放送技術研究所 NHK Science and Technology Research Laboratories

‡国立情報学研究所 National Institute of Informatics
総合研究大学院大学 The Graduate University of Advanced Studies

- **拡張固有表現による重み (w_{ene})**

統計的な情報によるスコアリングは S_{BM25} により行われているが、語の意味に基づいた関連性を考慮するため、拡張固有表現（ENE)[5]を利用した重み w_{ene} を付与する。ENE の抽出器は、独自作成の新聞及び EPG の ENE コーパスを用い、CRF++[1]を用いて作成した。素性には、各形態素の表層、品詞、EDR の概念 ID 等を用いた。今回の実験では、5 種類の ENE (PERSON, TITLE, ORGANIZATION, LIVING_THING, GPE)を使用し、それぞれの重みを予備実験により定めた。

- **n-gram による重み (w_{ng})**

本手法では、n=1, 2, 3 のすべての n-gram を素性に使用しているため、複合語に対して、過剰にスコアが加算される恐れがある。例えば、3-gram の「経済-産業-省」が共起すれば、その構成要素の 1-gram「経済」「産業」「省」や 2-gram「経済-産業」「経済-省」も共起するためである。そこで、各 n-gram に対して $w_{ng} = 1/n$ の重みを付与する。

4.2 関係ラベルの付与

検索結果とキーコンテンツとの関連性を示す情報（関係ラベル）を付与する。検索元コンテンツ Q があるときの番組データベース中のある番組 D のラベル L(D,Q)を(4)式で求める。拡張固有表現で重み付けされた S_{BM25} が最も高い n-gram をラベルとする。

$$L(D,Q) = \underset{T_n}{\mathrm{Argmax}}(S_{BM25}(T_n, D) \times w_{ene}(T_n)) \quad (4)$$

すなわち、検索元コンテンツ Q が異なれば、同一コンテンツ D であっても、関係ラベルは変化する場合がある。

5. 実験

5.1 関連番組検索システム

提案手法を用いて関連番組検索システムを試作した。検索対象の番組データベースには、3873 本（17 種類）の番組を用いた。これらの番組の EPG と番組 HP の紹介文を検索元・検索先の情報とした。CC はニュース番組の検索元の情報としてのみ利用した。関連番組の検索結果例を図 1 に示す。例は、ニュース 7 の 1 項目“中小企業のアジア進出支援”に関連する番組を検索した結果である。関連番組として、「クローズアップ現代」や「NHK スペシャル」が提示され、その関係ラベルとして、「中国」「経済産業省」「中小企業」など ENE や複合語が付与されている。試作システムでは、同一の関係ラベルを持つ番組をクラスタ（関連度の累積値の上位 3、その他）にまとめて提示している。

5.2 評価実験

視聴中の番組に関連する番組をどの程度検索できるかを調べるため、試作システムを用いた関連番組検索の評価実験を行った。実験では、種類に重複のない 15 番組を検索元とし、それぞれの検索結果の上位 10 番組を評価した。番組間の関連性を 4 段階（4:関連ある 3:少し関連ある 2:あまり関連ない 1:関連ない）で、関係ラベルを 3 段階（3:適切 2:ほぼ適切 1:不適切）で評価した。また、ベースラインとして、形態素の tf-idf 値を素性とした Cosine 類似度を用いた。ラベルには tf-idf 値が最も高い形態素を付与した。

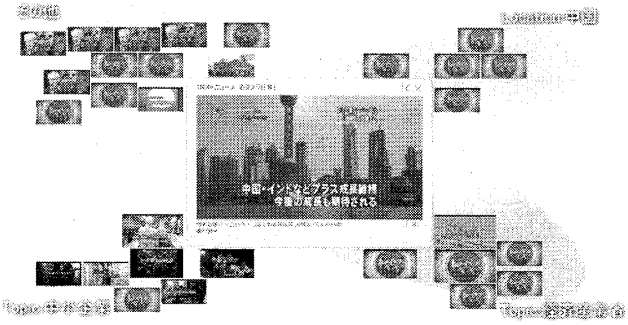

図 1　関連番組の検索結果例

表 1　実験結果

	関連性(Ave.)	関係ラベル(Ave.)
ベースライン	3.26	2.25
提案手法	3.35	2.55

表 1 に、それぞれの平均の評価値を示す。実験の結果、関連性と関係ラベルの評価ともに、提案手法がベースラインを若干上回った。提案手法は、検索元番組と共通の ENE や複合語を持つ番組の順位が上昇する傾向がある。そのため、今回の実験では、これらの番組が“関連ある”と判断され、評価結果が改善したと考えられる。一方で、少数ではあるが、番組関連テキスト内で例示に使われた国名など、内容とは直接関係ない ENE がスコアに寄与し、関連のない番組が提示される場合があった。誤った結果を削減するため、検索元コンテンツの文脈を考慮した選択的な ENE の重み付けを行うことが課題である。

関係ラベルの評価においても、複合語や ENE など意味のある語が関係ラベルに選択されたことで、ユーザの評価が改善したと考えられる。

6. まとめ

番組の内容を説明するテキストの類似度に基づく、視聴中の番組を起点とした関連番組の検索手法を提案した。今後、番組データベースの対象を拡大し、関連番組検索の性能評価を行う予定である。

参考文献

[1] 藤井真人, 柴田正啓, 住吉英樹, 後藤淳, 佐野雅規, 望月貴裕, 宮崎勝, 八木伸行, “情報検索を使う新しい視聴スタイル CurioVew －具体的化に向けた検討－, 2009 映情学年大 (2009).
[2] 溝口祐美子, 長野伸一, 稲葉真純, 川村隆浩, 中本利明, 浅川一満, “オントロジーを用いた TV 番組グラフ作成システム”, 信学技報、WI2-2007 (2007).
[3] 山崎智弘, 真鍋俊彦, 川村隆浩 “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, 25(4) pp.41-51 (2008).
[4] Stephen E. Robertson and Steve Walker, "Okapi/Keenbow at TREC-8" In Proceedings of the 8rd Text Retrieval Conference (1999).
[5] Satoshi Sekine, "Extended Named Entity Ontology with Attribute Information", In Proceedings of the 5th International Conference on Language Resources and Evaluation (2008).

[1] http://crfpp.sourceforge.net/