TV ニュース映像の話題の網羅性・一般性・受容度の可視化による視聴支援

Supporting TV News Reception by Visualizing the Contents Coverage, Generality, and Acceptance of the Topics

甲谷 優♥ 湯本 高行♦
小山 聡♠ 田島 敬史♠
田中 克己♠

Yutaka KABUTOYA Takayuki YUMOTO
Satoshi OYAMA Keishi TAJIMA
Katsumi TANAKA

TV のような受動的なメディアにおけるニュース報道は，新聞やインターネット等のメディアとは異なり，記事を取捨選択する等，ユーザとのインタラクションに乏しく，能動的に閲覧することができない．また，発信されるニュース情報や論評などが真に正しいものかどうか判断する材料も少ない．これらの問題点を解決するために，馬らは TV ニュース番組と同じ話題の Web ページ情報を同時に見せることで情報を補完/比較する手法を提案している．本研究では，Web ページによって TV ニュース番組の情報を補完できるかどうか，その妥当性を検証する．その上で，Web ページによる TV ニュースの情報補完のための 1 つの手法として，TV ニュース情報の，WWW 全体の情報と比較した上での話題の一般性，受容度を，音楽プレイヤにおけるスペクトラムアナライザの形で可視化する．これにより情報補完だけでなく，ユーザのメディアリテラシを補ったり，話題に対する興味を想起させる等，視聴支援を行う．

It is difficult to watch TV news in an active manner such that the user can interactively select TV news articles, because TV is originally a broadcast information media. It is also difficult for users to judge whether the information of TV News is valid because conventional TV contents are not directly linked with related or evidence information. One of the methods to cope with them is to provide complementary or comparative information of TV news obtained from other media such as Web etc. In this paper, at first, we examine to what extent there is Web information can complement against TV news articles in Web pages. Next, we propose a new way to complement TV by Web, called a "TV news spectrum analyzer", which visualizes the degrees of generality and social acceptance of TV news articles by using WWW. This system also supports to complement users' media literacy, and recollect user's interests.

♥ 学生会員 京都大学大学院情報学研究科修士課程
kabutoya@dl.kuis.kyoto-u.ac.jp
♦ 正会員 兵庫県立大学 大学院工学研究科
yumoto@eng.u-hyogo.ac.jp
♠ 正会員 京都大学大学院情報学研究科
{oyama, tajima, tanaka}@dl.kuis.kyoto-u.ac.jp

1. はじめに

Google は，Web 検索の結果を，PageRank [1] というランキングアルゴリズムによって順位付けを行っている．その検索結果順位は，Web ページを閲覧するユーザにとって，ページを取捨選択するための重大なヒントとなっている．それを示す根拠として，たとえば[2] によれば，ほとんどの検索の機会においてユーザは上位10 件中に含まれるページしか利用しないことがわかっている．ユーザはその上位10 件から1 つを選択し，リンクを辿る．このように，Web ページ閲覧においてはシステムとのインタラクションを通じてユーザは能動的にコンテンツを消費している．

一方で，マスメディアによる報道，とくにTV ニュース視聴においては，記事を取捨選択する等の能動的な視聴は困難である．そのためTV による放送はプッシュ型配信と呼ばれる．近年メディアリテラシと呼ばれるメディアを批判的に読み解く力が重要視されてきている．

これらの問題を解決するべく，あるTV 番組に対して，それと同じ話題を持つWeb ページにどのような情報が含まれるかを調べ[3], [4]，そのTV 番組の情報を補完・比較するという手法が提案されてきた[3], [4].

本研究では，Web ページでTV ニュース番組の情報をどの程度補完できるのか検証するために，まずTV ニュース番組とWeb ページに含まれる話題構造を比較した．ある話題について，「TV ニュース番組では言及されていないが特定のWebページなら言及されている」ような話題がなければ Web ページでTV ニュースを補完する意味はない．

次に本研究では，同じ主題について言及されているWeb ページの話題と比較することで，TV ニュース番組の話題の一般性，受容度を定量的に評価する．ここで，一般性とはその情報が如何に普遍的であるか，すなわちそこに含まれる話題がどのくらいの数のWeb ページで言及されているかを表し，受容度とはその情報が如何に支持されているか，すなわち同じ話題を言及するページがどのくらい人気があるのかを表す尺度であるとする．

これらを可視化する上で，本研究では無線機・送信機やPC 用の音楽プレイヤでよく見受けられるスペクトラムアナライザを用いる．図1 に一般的なスペクトラムアナライザで構成される2 次元グラフの例を示す．音波を対象としたものの場合，一般的には横軸に周波数，縦軸に音圧をとった2 次元のグラフが画面上に構成される．本提案手法におけるスペクトラムアナライザは，横軸に受容度，縦軸に一般性をとった2 次元グラフを構成する．

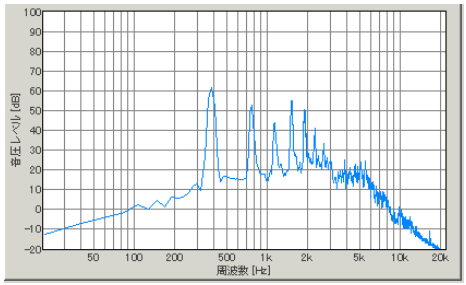

図1 スペクトラムアナライザの例
Fig.1 Example of 'Spctrum Analyzer'

本論文の構成は以下の通りである．まず2 章にて，本研究の関連研究について述べる．次に3 章にて，話題や網羅度，一般性や受容度等本研究の根幹である概念を定義する．4 章にてTVニュース番組とWeb ページの話題の網羅度を比較することによりWeb ページによるTV ニュース番組の補完の有効性を証明する．5 章ではスペクトラムアナライザの生成手順について述べる．6 章ではシステムの実装と評価実験，その考察について述べ，最後に7 章にてまとめと今後の課題について述べる．

2. 話題構造

本研究では，[4] にて Ma らの提案した話題構造・話題グラフ，話題構造の結合を利用している．本項ではそれらの定義を述べた後，それらをもとに新たに話題構造の評価尺度である話題構造の網羅度，一般性，受容度を定義する．さらに，文書から話題構造を抽出する手法を述べる．

2.1 話題構造

話題構造は以下に示すBNFで定義される[4].

$$
\begin{aligned}
\langle topic\rangle &::= (\{\langle subject\rangle\},\{\langle content\rangle\})\\
\langle subject\rangle &::= \langle subject\text{-}term\rangle\\
&\mid \langle subject\text{-}term\rangle,\langle subject\rangle\\
\langle content\rangle &::= \langle content\text{-}term\rangle\\
&\mid \langle content\text{-}term\rangle,\langle content\rangle\\
\langle subject\text{-}term\rangle &::= \langle keyword\rangle \mid \langle topic\rangle\\
\langle content\text{-}term\rangle &::= \langle keyword\rangle \mid \langle topic\rangle
\end{aligned}
\tag{1}
$$

すなわち話題構造とは，主題語集合・詳細語集合なる 2 つの語集合から構成される．

2.2 話題グラフ

話題グラフは話題構造に含まれるキーワードを節点で，2 つのキーワード間の主題語–詳細語関係を有向枝で表したものである．今，V を節点の集合，E を有向枝の集合とすると，話題構造 t の話題グラフは

$$G(t) = (V, E) \tag{2}$$

で定義される[5]．図2 の場合，V ，E はそれぞれ

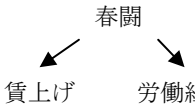

図 2　話題グラフの例
Fig.2　Example of a Topic Graph

$$V = \{\text{春闘}, \text{賃上げ}, \text{労働組合}\} \tag{3}$$

$$E = \{(\text{春闘}, \text{賃上げ}), (\text{春闘}, \text{労働組合})\} \tag{4}$$

となり，この場合の話題構造は以下のように示される．

$$t = (\{\text{春闘}\}, \{\text{賃上げ}, \text{労働組合}\}) \tag{5}$$

2.3 話題構造の結合

2 つの話題構造 t_1，t_2 の結合 $t_1 \quad t_2$ は，結合後の話題グラフが連結グラフとなるように定義される[4].

$$G(t_1 \bowtie t_2) = \begin{cases} (V_1 \cup V_2, E_1 \cup E_2), & \text{if } \kappa(G(t_1 \bowtie t_2)) \neq 0 \\ \phi, & \text{otherwise} \end{cases} \tag{6}$$

ただし，$\phi \times t = \phi$，$t \times \phi = \phi$ であり，また· はグラフの点連結度である．

話題構造の結合演算×が交換法則は満たすが結合法則は満たさないことは自明である．

2.4 話題構造の評価尺度

2.4.1 話題構造の網羅度

話題構造 t_1 の t_2 に対する網羅度を以下の式で定義する．

$$\mathrm{cov}(t_1, t_2) = \frac{|V_1 \cap V_2|}{|V_2|} \tag{7}$$

ただし

$$G(t_1) = (V_1, E_1),\ G(t_2) = (V_2, E_2) \tag{8}$$

また，以下を満たす t_1，t_2 が存在するのは自明である．

$$\mathrm{cov}(t_1, t_2) \neq \mathrm{cov}(t_2, t_1) \tag{9}$$

ここで，$\mathrm{cov}(t_1, t_2) = 1$ になるとき，t_1 は t_2 を包含するとし，$t_2 \subset t_1$ で表すことにする．

2.4.2 話題構造の一般性

話題構造の一般性とは，その話題構造が如何に多くの文書で出現するか，如何に普遍的であるかを示す尺度とする．したがって，話題構造の一般性は，その話題構造を含む文書数で定義できる．

2.4.3 話題構造の受容度

話題構造の受容度とは，その話題構造が如何に多くの人に支持されているかを示す尺度であるとする．本研究では，この尺度を Google の検索結果順位を用いて近似する．その根拠として，Google の検索順位は PageRank という被リンク数が多ければ多いほどスコアの高くなるランキングアルゴリズムによって決まっているからである．「リンクする」という行為をユーザの人気投票だと考えれば，話題構造の受容度はそれを含むページの PageRank 値，すなわち Google ウェブ検索での順位に帰着する．

2.5 話題構造の抽出

本研究では，[4] と同様，段落が 1 つの話題構造を持つと仮説を置いている．まず，語の共起度について定義し，さらにそこから語の主題語度・詳細語度を定義する．そして，最後に話題構造の抽出ステップについて言及する．

2.5.1 共起度

2 つのキーワード集合 W_i と W_j の無向共起度を，以下の式で定義する．

$$cooc(W_i, W_j) = \frac{df(W_i \cup W_j)}{df(W_i) + df(W_j) - df(W_i \cup W_j)} \tag{10}$$

ただし，$df(W)$ はキーワード集合 W 内の全てのキーワードの論理積をクエリとしたときの Google ウェブ検索結果のヒット数を表す．

また，2 つのキーワード集合 W_i と W_j の有向共起度を，以下の式で定義する．

$$\overrightarrow{cooc}(W_i, W_j) = df(W_i \cup W_j) / df(W_i) \tag{11}$$

2.5.2 主題語度，詳細語度

あるキーワード w_i のある話題構造 t 中における主題語度を，その話題構造と対応する文書 d 中の tf/idf 値[5] $tfidf(d, w_i)$ と t に含まれる他の語への有向共起度から，以下のように定義する．

$$sub(t, w_i) = tfidf(p, w_i) + \sum_{w_j \in t - w_i} \overrightarrow{cooc}(w_i, w_j) \tag{12}$$

また，あるキーワード w_i のある話題構造 t 中における詳細語度は，その話題構造に含まれる主題語集合の各キーワードとの無向共起度から以下のように定義される[4].

$$con(t, w_i) = \sum_{w_j \in S_i} cooc(w_i, w_j) \tag{13}$$

2.5.3 話題構造の抽出手順

ある文書 d が与えられたとき, 以下の手順から話題構造を抽出する.

1. d に含まれるテキストを Sen[1]を用いて形態素解析
2. 形態素のうち名詞のみを抽出
3. tfidf 法を用いて各キーワードの特徴量を計算，その値の高い何個かで話題構造の主題語集合を構成
4. それぞれの主題語に対し，含まれなかった語の詳細語度を計算，そこからいくつか話題構造を抽出
5. 先の手順にて抽出された話題構造を結合し，極大なものを得る

3. TV ニュースの一般性・受容度の可視化

本研究では，TV ニュース番組のクローズドキャプション中の話題構造を抽出し，それを基に品質評価を行う．その品質をスペクトラムアナライザに表示される2次元グラフで可視化する．本項ではまず，一般的なスペクトラムアナライザに表示されるグラフについて説明し，次に本研究で提案するスペクトラムアナライザの持つ情報について説明する．最後にその作成手順について述べる．

3.1 一般的なスペクトラムアナライザ

スペクトラムアナライザとは，本来電気計測器であり，ある時刻における電波の周波数を横軸に，電力及び電圧を縦軸とする 2 次元グラフを画面に表示するものである．近年は PC 用の音楽プレイヤソフトや，一部のラジオカセットレコーダ，カーオーディオにもこのスペクトラムアナライザの機能がある. この場合は音波の周波数と音圧により 2 次元グラフを構成する.

3.2 可視化したグラフの提供する情報

本研究で提案するスペクトラムアナライザは，TV ニュース番組に含まれる話題の一般性と受容度という 2 つの側面を可視化する.

3.3 スペクトラムアナライザの作成手順

以下のような手順でクローズドキャプションからスペクトラムアナライザを作成する.

1. クローズドキャプションをニュース単位で分割
2. 分割後の各セグメントから話題構造を 1 つ抽出
3. それぞれの話題構造から主題語集合を抽出，その論理積をクエリとして Google ウェブ検索
4. 検索結果上位 100 件を順位でクラスタリング
5. 各順位クラスタの話題を抽出，クローズドキャプションの話題と比較

以下，それぞれの手順の詳細を述べる.

3.3.1 クローズドキャプションの分割粒度

このクローズドキャプションと映像には

- 映像全体，プログラム
- シーン
- カット(ショット)
- 文
- 語(形態素)

のように階層構造があり，どの粒度を単位として話題を抽出するかで作成されるスペクトラムアナライザも異なってくる．ただし，語と(話題抽出には少なくとも 2 語以上必要), 映像全体(そこに含まれる全ての話題を結合できず, また話題が変化しなくなってしまう) の粒度では単一の話題を抽出できない．そこで本研究ではシーンを単位として話題構造を抽出し，それをもとにスペクトラムアナライザを作成する.

3.3.2 検索順位によるクラスタリング

本研究では, 大規模ネットワークのもつスケールフリー性[6]に基づき，Google ウェブ検索の結果をそれらの順位にしたがっていくつかの順位クラスタに分割する．大規模ネットワークのもつスケールフリー性とは，それにおいて

- 特徴的なスケールが存在しない
- 分布が著しく非対称

という性質のことである．たとえば，PageRank のスコアはべき分布していることが知られている．したがって，Google 検索結果 1 位と 2 位の差と 100 位と 101 位の差は全く異なる．この性質に基づき，スペクトラムアナライザを作成するのに n 件の検索結果を用いて p 個の順位クラスタを作成する場合, i 番目のクラスタ WEB_i には $\lfloor \alpha^{i-1} \rfloor$ 位から $\lfloor \alpha^{i} \rfloor$ 位のものを属させる．ただし

$$\alpha = \sqrt[p]{n} \tag{14}$$

3.4 生成されるスペクトラムアナライザ

1 つの TV ニュース番組がシーン $scene_j\ (j = 1,2,\ldots)$ に分割されたとき, その分割後のセグメント $scene_j$ に対しスペクトラムアナライザ sp_j は以下のように生成される.

$$sp_j = (st_j, et_j, val_{j1}, val_{j2}, val_{j3}, val_{j4}) \tag{15}$$

st_j, et_j はそれぞれ $scene_j$ の始まる時間と終わる時間である. これらの情報はクローズドキャプション内の時間タグから取得可能である．val_{ji} はスペクトラムアナライザの各バーの高さを表しており，ウェブ検索結果のクラスタ WEB_{ji} から計算される.

$scene_j$ から抽出された話題構造を t_j とする．(1) より

$$t_j = (S_j, C_j) \tag{16}$$

ただし S_j は主題語集合，C_j は詳細語集合である.

ここで，$scene_j$ と WEB_{ji} の両者を話題構造 t_j に基づき特徴ベクトル化し，両者のコサイン相関値をとることでそれを val_{ji} とする．特徴ベクトルの各次元は各詳細語 $w_k \in C_j$ に対応している.

3.4.1 $scene_j$ の特徴ベクトル化

$scene_j$ を特徴ベクトル化したものを $\boldsymbol{s}_j$ とし，$\boldsymbol{s}_j$ の w_k に対応する要素の値を，そのキーワードの，話題 t_j の核 w_i の出現する文書の中での IDF 値とする．たとえば $\boldsymbol{s}_j$ の $w_k \in C_j$ に対応する要素を

$$(e_k, s_j) = \frac{1}{\overrightarrow{cooc}(S_j, \{w_k\})} \tag{17}$$

と定義する. ただし, e_k は $w_k \in C_j$ に対応する単位ベクトル.

3.4.2 WEB_{ji} の特徴ベクトル化

$scene_j$ とほぼ同様の考え方で順位クラスタ WEB_{ji} を特徴ベクトル化し web_{ji} を得る.

まず WEB_{ji} のなかで単語集合 W の全てを含む文書数を $df_{ji}(W)$ とする.

このとき，$\boldsymbol{web}_{ji}$ の w_k に対応する要素を

$$(e_k, web_{ji}) = \frac{df_{ji}(S_j \cup \{w_k\})}{\overrightarrow{cooc}(S_j, \{w_k\})} \tag{18}$$

と定義する.

3.4.3 val_{ji} の計算

したがって，val_{ji} は $scene_j$, WEB_{ji} の特徴ベクトル s_j, web_{ji} を用いて以下の式から計算される.

$$val_{ji} = \frac{(s_j, web_{ji})}{|s_j||web_{ji}|} \tag{19}$$

[1] http://ultimania.org/sen/

ただし，$|\boldsymbol{v}|$ は $\boldsymbol{v}$ のユークリッドノルムである．

3.5 システムの実装

図 3 にシステムの外観を示す．下部の赤い楕円で囲まれたスペクトラムアナライザが時間とともに変化する．

上位クラスタに対応するバー(より左側のバー)の高さが高いほど受容度が高いということを表している．なぜなら，上位クラスタ内のページに類似した話題構造を持つページが多いということは，その話題構造は人気度の高いページに多く存在するということであり，多くの人に支持されているということになると考えられるからである．

また，順位クラスタに属するページの件数は順位が高いほど少ない．つまりより右のバーの高さが高いほど，現在視聴中のニュース内の話題と類似した話題を持つページが多いということを表す．それはすなわち，そのニュースの話題が一般的であるということを表す．

勿論，右のバーの高さも左のバーの高さも高いときは，そのニュースに含まれる話題の一般性・受容度がともに高いということを表している．

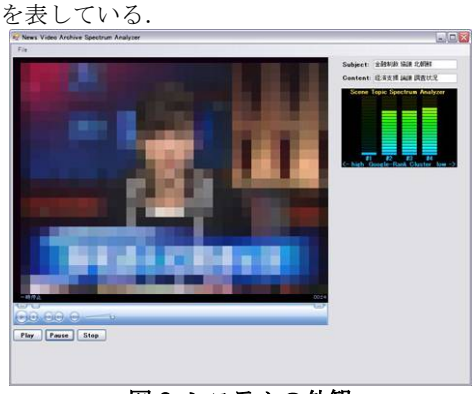

図 3 システムの外観
Fig.3 System Overview

いくつかのニュースで実験したところ，一般性及び受容度のうち両方とも低いようなものは存在しないことがわかった．

4. まとめと今後の課題

本研究ではTV ニュース番組の情報の一般性・受容度をWebページと比較し，可視化する手法を提案した．

TV ニュース番組に含まれる話題構造をWeb ページと比較することによりその情報の一般性と受容度を可視化することで視聴支援を行う手法を提案した．

今後の課題としては，品質に関するより詳細な情報の可視化を考えている．具体的には，比較した Web ページのスニペットを集約して補完情報として与える手法，比較した TV ニュース及び Web ページの話題構造をグラフとして可視化する手法が考えられる．

比較 Web ページを獲得するために，今回の研究では TV ニュースのクローズドキャプションからクエリを作成した．今後の課題として，この手法を発展させていくことを考えている．具体的には，ある Web ページが与えられたとき，それを検索結果としてヒットさせる最も典型的なクエリを作成する手法の提案，検索エンジンを通じた関連文書検索手法の提案を考えている．

[謝辞]

本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」(リーダー：田中克己，平成 14～18 年度)，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発(研究代表者：田中克己)，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者：田中克己，A01-00-02，課題番号 18049041)，および文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」(研究代表者：安達淳，Y00-01，課題番号：18049073) によるものです．ここに記して謝意を表すものとします．

[文献]

[1] L. Page, S. Brin, R. Motwani and T. Winograd: “The pagerank citation ranking: Bringing order to the web” (1998).

[2] T. Joachims: “Optimizing search engines using clickthrough data”, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 133–142 (2002).

[3] M. Henzinger, B. Chang, B. Milch and S. Brin: “Query-Free News Search”, World Wide Web, 8, 2, pp. 101–126 (2005).

[4] Q. Ma and K. Tanaka: “Topic-Structure Based Complementary Information Retrieval for Information Augmentation”, Lecture Notes in Computer Science (APWeb2004), pp. 608–619 (2004).

[5] G. Salton: “Automatic Information Organization and Retrieval.”, McGraw Hill Text (1968).

[6] A.L., R. Albert : “Emergence of Scaling in Random Networks” , Science, 286, 5439, p. 509 (1999).

甲谷 優 Yutaka KABUTOYA
京都大学大学院情報学研究科修士課程在学中．日本データベース学会学生会員．

湯本 高行 Takayuki YUMOTO
兵庫県立大学大学院工学研究科助教．2007 京都大学大学院情報学研究科博士後期課程修了．博士(情報学)．主に情報検索・情報統合に関する研究に従事．情報処理学会，日本データベース学会，ACM，IEEE 各会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助教．2002 年京都大学大学院情報学研究科博士後期課程修了．博士(情報学)．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田島 敬史 Keishi TAJIMA
京都大学大学院情報学研究科社会情報学専攻准教授．1996 年東京大学理学系研究科情報科学専攻博士後期課程修了．博士(理学)．主にデータベースプログラミング言語，Web 検索の研究に従事．IEEE Computer Society，ACM，情報処理学会，日本データベース学会等各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程終了．博士(工学)．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．