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

甲谷　優[†]　湯本　高行[†]　小山　聡[†]　田島　敬史[†]　田中　克己[†]

† 京都大学大学院情報学研究科 〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{kabutoya,yumoto,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　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．
キーワード　可視化，話題構造，網羅度，一般性，受容度

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

Yutaka KABUTOYA[†], Takayuki YUMOTO[†], Satoshi OYAMA[†], Keishi TAJIMA[†], and Katsumi TANAKA[†]

† Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan
E-mail: †{kabutoya,yumoto,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　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.
Key words　Visualization, Topic Structure, Contents Coverage, Generality, Social Acceptance

1. はじめに

Google(注1) 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．

(注1)：http://www.google.co.jp/

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], [5]，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], [5].

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．

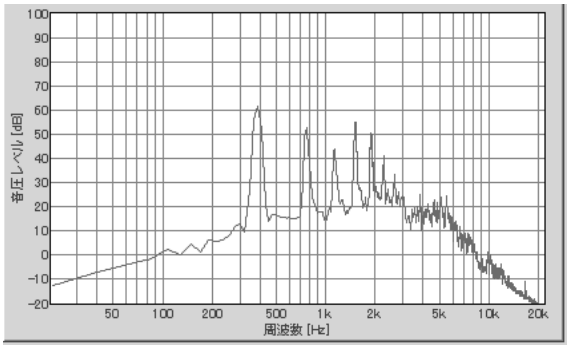


図 1　スペクトラムアナライザの例

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. 関 連 研 究

2.1 話題構造抽出

大澤らはある文章がどのような内容であるか，どのような話題構造を持っているかを調べるために，重要であろうと考えられる語を複数抽出し，共起するものどうしを線で結ぶことによりグラフ表示し可視化する，KeyGraph [4] と呼ぶ手法を提案している．

また，Ma らは文書から話題中心となる主題語とそれ以外の詳細語からなる話題構造を抽出する手法を提案している [5]．本研究では [5] の手法を用いて文書から話題構造を抽出し，それをもとに 2 つの文書を比較している．

2.2 リンクを持たないコンテンツの品質推定

筆者らはこれまでに，Web 検索結果における順位情報と，類似度を用いてリンクのないコンテンツに対するランキングを行ってきた [6]．PageRank は Web ページ間に存在するリンクという関係を人気投票のように捉え，Web ページの品質を人気度という側面から評価するアルゴリズムだが，リンク構造を持たないコンテンツ，たとえば HD レコーダ内の録画番組には適用できない．そこで，類似しているコンテンツは人気度も近いと仮説を立て，それらリンクのないコンテンツの人気度を測定した．

この手法の問題点，今後の課題として，主に以下に示す 2 点が存在する．

（1） 対象がリンクを持たないコンテンツと曖昧であり，実際にどのような検索を行う際にこの手法が活かされるのかわからない．

（2） ランキングが必要となる検索は，検索結果の総数が膨大であるときである．人気度の測定できないようなものが対象となる検索で，ランキングが必要となるほど膨大なデータが存在することは現在のところない．

すなわち，具体的なシステムが想定されていないことが最大の問題点であった．本研究を筆者らは，その品質評価の 1 つの具体的な応用として位置付けている．

一方で，Kurland らは言語モデルからリンクのないコンテンツ間に存在する仮想的なリンクを発見し，それを基に PageRank を計算する手法を提案している [7]．この研究は，品質推定よりもコンテンツ間に存在する潜在的なリンクを発見することに重きを置いており，我々の研究とは異なる．

3. 話 題 構 造

本研究では，[5] 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．

3.1 話 題 構 造

話題構造は以下に示す BNF で定義される [5].

$$\langle topic \rangle ::= (\{\langle subject \rangle\}, \{\langle content \rangle\})$$

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

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．

3.2 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

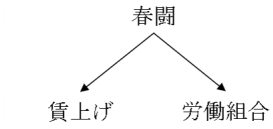


図 2　話題グラフの例

$$V = \{\ 春闘, 賃上げ, 労働組合\ \} \tag{3}$$

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

となり，この場合の話題構造は以下のように示される．

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

3.3 話題構造の結合

2 つの話題構造 t_1，t_2 の結合 $t_1 \bowtie t_2$ 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 [5].

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

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

話題構造の結合演算 $\bowtie$ 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．

3.4 話題構造の評価尺度

3.4.1 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．

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

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．

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

ここで，$cov(t_1, t_2) = 1$ になるとき，t_1 は t_2 を包含するといい，$t_2 \prec t_1$ で表す．

3.4.2 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．

3.4.3 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．

3.5 話題構造の抽出

本研究では, [5] 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．

3.5.1 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．

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

3.5.2 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段落 p 中の tf/idf 値 [8] $tfidf(p, w_i)$ 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 [5].

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

3.5.3 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） 文書 d に含まれるテキストを段落毎に分割 $(p_1, p_2, \ldots)$

（2） $p_1, p_2, \ldots$ を Sen [注2] を用いて形態素解析

（3） 形態素のうち名詞のみを抽出

（4） 連続して出現する名詞のうち品詞の詳細によって接続させて名詞句とし，これらをキーワードとする．名詞句を表す BNF の一部を以下に示す．

$$
\begin{aligned}
\langle \text{名詞句} \rangle ::= & \langle \text{接続名詞} \rangle \langle \text{名詞句} \rangle \\
| & \langle \text{名詞句} \rangle \langle \text{接続名詞} \rangle \\
| & \langle \text{左辺接続名詞} \rangle \langle \text{名詞句} \rangle \\
| & \langle \text{名詞句} \rangle \langle \text{右辺接続名詞} \rangle \\
\langle \text{接続名詞} \rangle ::= & \langle \text{一般名詞} \rangle | \langle \text{固有名詞} \rangle \\
\langle \text{左辺接続名詞} \rangle ::= & \langle \text{接頭詞} \rangle \\
\langle \text{右辺接続名詞} \rangle ::= & \langle \text{接尾名詞} \rangle
\end{aligned} \tag{14}
$$

（5） 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

（6） 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

（7） 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

4. TV ニュースと Web ページの話題構造比較

本章では，TV ニュースと Web ページという異なる 2 つのメディアに含まれる話題構造を比較することにより，両者に内容の差があるのかを比較する．

4.1 実験 1: TV ニュース番組の WWW に対する話題の網羅度

これまでにいくつかの TV ニュース番組の情報を Web ページで補完する手法が提案されてきている．しかし，TV ニュース番組に含まれる話題がすべての Web ページの話題を包含する場合，Web ページで TV ニュース番組の情報を補完することはできない．したがって，本項では TV ニュースに含まれる話題が WWW 内の話題を包含しないことを証明するために，両者の話題構造を比較する実験を行う．

4.1.1 実験データ

本研究では TV ニュース番組のクローズドキャプションを使用する．

```
[029] 19:05:26 00103
[030] 19:05:28 00105 安倍総理大臣は
[031] 19:05:30 00107 政権発足後、
[032] 19:05:32 00109 初めての国会となった
[033] 19:05:35 00112 第１６５臨時国会が閉会した
[034] 19:05:37 00114 改正教育基本法など
[035] 19:05:40 00117 重要法案が
[036] 19:05:42 00119 成立し、
```

図 3 クローズドキャプションの例

図 3 のように，番組中の音声データに時間タグをつけたテキストストリームがクローズドキャプションである．

次に TV ニュース番組として，日本放送協会 (NHK) によって提供されている「ニュース 7」の 2006 年 12 月 19 日 19 時 00 分〜19 時 30 分放送分のうち，第 165 回臨時国会閉会に関するシーン (19 時 05 分〜19 時 8 分) を用意した．以下，そのシーンの要約である．

> 安倍総理大臣は政権発足後，初めての国会となった第 165 臨時国会が閉会したのを受けて記者会見した．改正教育基本法等重要法案が成立し，戦後体制から脱却して新たな国づくりを行う礎ができたと成果を強調した．

次に WWW として，Google に対して適当なクエリを実行したときの検索結果ページ上位 20 件を用いる．

4.1.2 実 験 手 順

以下の手順より実験を行う．

（1） 実験データのシーンに対応するクローズドキャプションから話題構造 t_{tv} を抽出

（2） 話題構造 t_{tv} の主題語集合を S_{tv} とする

（3） S_{tv} の論理積をクエリとして Google ウェブ検索し，上位 20 件のページ $p_i (i = 1, 2, \ldots)$ を獲得

（4） それぞれの p_i に対して

（a） パラグラフ毎に分割，その中から S_{tv} を含むものを選択

（b） そのパラグラフに対応する話題構造 $t_{\mathrm{web}i}$ を抽出

（5） 得られた話題構造を結合し，話題構造 t_{web} を得る

$$
t_{\mathrm{web}} = t_{\mathrm{web1}} \bowtie t_{\mathrm{web2}} \bowtie \ldots \tag{15}
$$

（6） $cov(t_{\mathrm{tv}}, t_{\mathrm{web}})$ を計算

4.1.3 結　　果

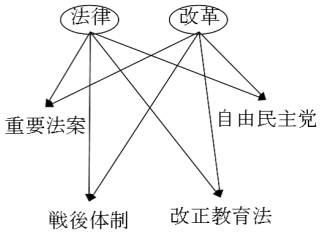


図 4 TV ニュース映像に含まれる話題構造 t_{tv}

図 4 に TV ニュース映像に含まれている話題構造 t_{tv} を示す．主題語が法律，改革の 2 語であり，詳細語が重要法案，戦後体制，改正教育法，自由民主党の 4 語である．この話題構造の主題語である法律，改革の 2 語の論理積をクエリとして Google ウェブ検索をする．以下がその 検索結果上位 20 件のうちの一部のタイトルである．

1. 中央省庁等改革関連法律
2. 司法制度改革関連法案
5. 第 164 回 通常国会 - 内閣府
10. 行政改革について：行政改革推進事務局ホームページ
11. 連合｜法律・制度改革 (分類別インデックス)

(注2)：http://ultimania.org/sen/

13. 公益法人制度改革 - Wikipedia
15. 選挙制度改革

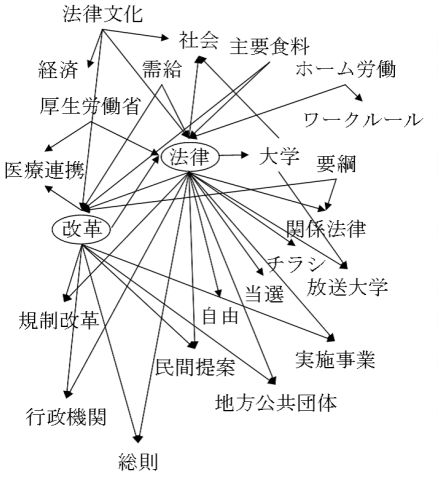


図 5 WWW に含まれる話題構造 t_{web}

これら 20 件から関連するパラグラフを選択，そこから話題構造を抽出して結合したもの t_{web} を図 5 に示す．証明すべきはこの話題構造 t_{web} が図 4 に示した話題構造 t_{tv} に包含されないこと，すなわち $cov(t_{\mathrm{tv}}, t_{\mathrm{web}}) < 1$ が成立することである．

まず，t_{web} に含まれるキーワードの数が 25 であり，また t_{tv} と t_{web} の両方に含まれるキーワードの数が 2 であるので，

$$\begin{aligned} cov(t_{\mathrm{tv}}, t_{\mathrm{web}}) &= 2/25 \\ &= 0.08 < 1 \end{aligned} \tag{16}$$

したがって $t_{\mathrm{web}} \not\prec t_{\mathrm{tv}}$.

4.2 実験 2: TV ニュースと Web ニュースの話題構造比較

ニュース情報は現在 TV や Web，ラジオ，新聞といったさまざまなメディアを介して配信されている．本研究ではその中でも特に Web と TV という 2 つの異なるメディアに着目し，それらの話題構造を比較する．それにより両者の内容に差があるのか，またあればそれがどのようなものかを検証する．

4.2.1 実験データ

TV ニュース 番組は先の 4.1 節のものと同じものを使用する．

Web ニュース 記事は，手動で同時期の同内容のものをニュースサイトの MSN 毎日インタラクティブ(注3)から 1 つ選択し利用する．以下がそのタイトルと要約である．

> 臨時国会：継続案件を優先，政府提出の全法案成立，きょう閉幕
>
> 臨時国会は既に法案審議を終え，１９日に閉幕する．８５日間の会期は００年以降の臨時国会としては最長だったが，安倍晋三首相が最重要法案に掲げた改正教育基本法など前国会からの継続案件の審議を優先している．

(注3)：http://www.mainichi-msn.co.jp/

4.2.2 実験手順

手順も 4.1 節のときのものとほぼ同様である．

（1） 実験データのシーンに対応するクローズドキャプションから話題構造 t_{tv} を抽出
（2） 同様に Web ニュース記事から話題構造 t_{news} を抽出
（3） t_{tv} と t_{news} を比較

4.2.3 結果

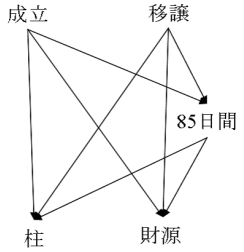


図 6 Web ニュース記事に含まれる話題構造 t_{news}

図 6 に Web ニュース記事に含まれる話題構造 t_{news} を示す．t_{tv} と比較すると両者ともに含まれるキーワードの総数は 0 個であるので

$$cov(t_{\mathrm{tv}}, t_{\mathrm{news}}) = cov(t_{\mathrm{news}}, t_{\mathrm{tv}}) = 0 \tag{17}$$

が成立する．このことはすなわち同じ時期の同じ事象に関する記事であるにもかかわらず，内容的には両者がまるで似ていないという結論に至る．

4.3 考察

前半の実験により，一般の Web ページから TV ニュース映像に対する十分な補完情報を得ることができることがわかった．たとえば 4.1 節の実験例の場合，改正教育法等に関して審議された第 165 回臨時国会の閉会に対する安倍総理のコメントに関する映像に対して，具体的な改正教育法の内容に関するページや，今年度実施された行政改革の内容に関するページで補完できる．

後半の実験結果からは，たとえ同じ時期の同じ事象に関するニュース記事であっても，TV と Web とでは内容に差があることがわかった．その原因としては，

- トピックの構造化・細分化に関して
  - TV ニュースの場合トピックの構造化は困難であり，また現在のシーンがどういうカテゴリのものなのか (政治や経済等) を伝えることができない．
  - Web ニュースの場合，内容・トピックを構造化・細分化することができる．たとえば今回扱った Web ニュース記事は政治/国会欄にあった記事だが，政治/行政欄にも同じトピックに関する別の記者の記事が存在した．
- 音声・映像による配信と活字・テキストによる配信の差
  - 音声・映像による配信の場合，視聴者に誤解を与えてはならないため，表現を簡略化することができない．たとえば自由民主党を自民党と表現できない．また音声を字幕データに落とすときに入力ミスが発生する．

等考えられる．また TV ニュースは Web のものと比較してメディアリッチであるという自明な特性がある．他にも，一方は

NHK によるもので，一方は毎日によるものであったため，両者の社説や主眼が異なったことも考えられる．

5. スペクトラムアナライザ

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，スペクトラムアナライザを作成する．まず，一般的なスペクトラムアナライザについて説明し，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．

5.1 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．

5.2 作成したスペクトラムアナライザの提供する情報

4 章にて証明したように，TV ニュース番組では言及されてないが Web ページでは言及されている話題が存在する．一方で逆に TV ニュース番組で言及されている話題が全ての Web ページで言及されているとは限らない．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．

5.3 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ある粒度で分割
（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

図 7 に手順のおおまかな流れを示す．

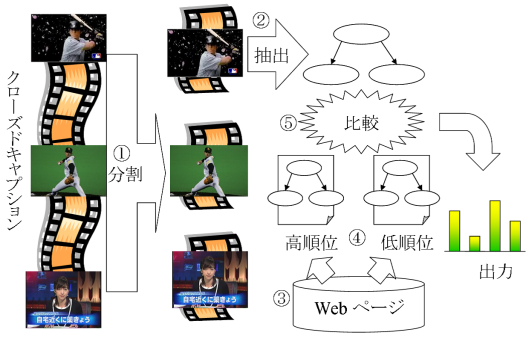


図 7 スペクトラムアナライザの作成手順

以下，それぞれの手順の詳細を述べる．

5.3.1 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．

5.3.2 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 [9] 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 $\lceil \alpha^{i-1} \rceil$ 位から $\lfloor \alpha^i \rfloor$ 位のものを属させる．ただし

$$\alpha = \sqrt[p]{n} \tag{18}$$

5.4 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．

$$sp_j = (st_j, et_j, val_{j1}, val_{j2}, val_{j3}, val_{j4}) \tag{19}$$

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

$$t_j = (S_j, C_j) \tag{20}$$

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．

5.4.1 $scene_j$ 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

$$(\boldsymbol{e}_k,\ \boldsymbol{s}_j) = \frac{1}{\overrightarrow{cooc}(S_j, \{w_k\})} \tag{21}$$

と定義する．ただし，$\boldsymbol{e}_k$ は $w_k \in C_j$ に対応する単位ベクトル．

5.4.2 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 $\boldsymbol{web}_{ji}$ 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

$$(\boldsymbol{e}_k,\ \boldsymbol{web}_{ji}) = \frac{df_{ji}(S_j \cup \{w_k\})}{\overrightarrow{cooc}(S_j, \{w_k\})} \tag{22}$$

と定義する．

5.4.3 val_{ji} の計算

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

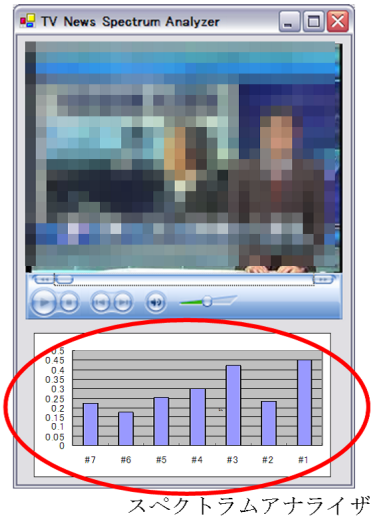
$$val_{ji} = \frac{(\boldsymbol{s}_j, \boldsymbol{web}_{ji})}{|\boldsymbol{s}_j||\boldsymbol{web}_{ji}|} \tag{23}$$

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

6. システムの実装

6.1 システムの外観

図 8 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．各バーの高さが高いほどその時点での話題構造の一般性が高いということを示し，また上位クラスタに対応する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，いずれの順位クラスタ内のいずれのページもその映像と類似した話題構造を持っているならばその話題構造は普遍的であるということであり，また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からである．



スペクトラムアナライザ

図 8　システムの外観

6.2 実験 3: スペクトラムアナライザの作成

本システムがどのように動作するか検証するため，実際の映像からスペクトラムアナライザを作成する実験を行った．

6.2.1 実験データと実験手順

本実験では日本放送協会 (NHK) の提供する番組「ニュース 7」の 2006 年 12 月 19 日放送分と，そこから抽出された話題構造の主題語をクエリとして Google ウェブ検索を実行した際の上位 100 件を対象として実験を行った．実験の手順は以下に示す通りである．

表 1　2006 年 12 月 19 日「ニュース 7」分割後のシーンとその内容

シーン	内容
$scene_2$	第 165 臨時国会閉会時の安倍総理の記者会見
$scene_3$	日本経団連による来年の春闘の基本方針
$scene_6$	ノロウィルスの感染予防，カキへの経済的被害
$scene_7$	打越さん，六甲山から生還
$scene_8$	NHK 紅白歌合戦出場者決まる
$scene_9$	クマの捕獲数，クマによる被害が過去最悪
$scene_{10}$	村上ファンドのインサイダー取引事件，宮内元取締役の証言
$scene_{12}$	麻生外務大臣，新たな派閥「為公会」結成
$scene_{16}$	明日の天気予報，東京，沖縄，...

（1） ニュース映像を手動により 16 シーン $scene_j (j = 1, 2, \ldots, 16)$ に分割．分割後のシーンとその内容を表 1 に示す．

（2） 以下，$j = 1, 2, \ldots, 16$ に対し，

（a） それぞれのシーンから話題構造 t_j を抽出．

（b） それぞれの話題構造の主題語により Google ウェブ検索，上位 100 件を獲得．

（c） ネットワークの持つスケールフリー性に基づき検索結果を順位にしたがって 4 つの順位クラスタ $WEB_{ji} (i = 1, 2, \ldots, 4)$ に分割．

（d） 式 (21)，(22) により $scene_j$，WEB_{ji} の特徴ベクトル $\boldsymbol{s}_j$，$\boldsymbol{web}_{ji}$ を計算．

（e） 式 (23) により val_{ji} を計算．

6.2.2 実験結果

表 2　スペクトラムアナライザの作成例

シーン	val_{j1}	val_{j2}	val_{j3}	val_{j4}
$scene_2$	0.652	0.775	0.460	0.520
$scene_3$	1.000	0.994	0.991	0.999
$scene_6$	0.787	0.027	0.889	0.843
$scene_7$	0.425	0.434	0.562	0.786
$scene_8$	0.000	0.040	0.860	0.992
$scene_9$	0.583	0.599	0.666	0.733
$scene_{10}$	0.999	1.000	0.997	0.996
$scene_{12}$	0.397	0.758	0.793	0.985
$scene_{16}$	0.000	0.151	0.925	0.976

実験結果を表 2 に示す．val の値が高いほど，そのシーンの話題構造の一般性が高いことを表しており，val の添え字の 2 文字目が若いほどそのシーンの話題構造の受容度に与える影響が大きいことを表している．この表から，シーン毎に話題構造の一般性，受容度に関していくつかのパターンが存在することがわかる．

- **一般性，受容度ともに大であるもの**
  全ての順位クラスタに対して話題構造の類似度が大，すなわちどの順位クラスタ内のいずれのページもそのシーンと類似した内容になっているようなシーンが存在した．このようなシーンは一般性，受容度ともに大であると言える．$scene_3$，$scene_{10}$ がこの場合にあてはまる．
- **一般性は大だが，受容度はさほど大きくないもの**
  高順位のクラスタ内のページには類似した話題構造をもつものはないが，低順位クラスタ内のページには類似した話題構造を持つものが多くあるシーンが存在した．このシーンは一般性は大であるが受容度はさほど大きくはないと言える．$scene_8$，$scene_{16}$ がこれにあてはまる．
- **一般性，受容度ともにさほど大きくないもの**
  全ての順位クラスタに対して似た話題構造を持つページも似ていない話題構造を持つページも存在するようなシーンが存在した．このシーンの一般性・受容度はともにさほど大きくはないと言える．$scene_2$ や $scene_7$ がこれにあてはまる．

6.2.3 考　　察

扱った番組がニュースということもあり，情報が洗練されているためか，総じて一般性・受容度ともに高かった．これが例えばバラエティ番組を対象としたときの値がどうなるか，我々の興味に尽きない．また，大半のシーンでスペクトラムアナライザがおしなべて高い値を計時する一方で，一般性は高いものの，受容度の低いシーンが存在することもわかった．この現象と，それが生じるシーンの内容を照らし合わせて鑑みると，話題構造のいわば『俗っぽさ』のようなものが評価できている可能性があることがわかる．

7. 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．また，Web と TV という 2 つの異なるメディアを話題構造の網羅度という側面から比較した．

7.1 話題構造からの TV と Web のメディア比較

本研究では TV ニュース番組と一般の Web ページ，Web ニュース記事を話題構造を比較することにより TV ニュース番組の情報を Web で補完できることを証明した．また，4.2 節の実験からは， TV ニュース番組と Web ニュース記事の間にメディアリッチであるか否かという特性以外の違いが存在することがわかった．TV ニュース番組は Web ニュースと比較して内容のカテゴリを明示できず，また音声・映像による配信であるために明瞭な表現を使わなければならないという制約を受けるため，内容に若干の差が生じる．

今回の実験では 1 例のみの結果による考察であったために，もう少し実験を続けて結果例を増やしていく必要がある．また，4.2 節の実験については，TV ニュースと Web ニュースの内容の差は社説の違いに過ぎない可能性があるために，実験データを変えて (たとえば毎日テレビのニュース番組と MSN 毎日インタラクティブの記事) 再実験してみる必要がある．

7.2 TV ニュース番組に含まれる話題のスペクトル分析

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．本研究で提案した一般性・受容度は互いに直交する概念であるので，本提案手法は一種のスペクトル分析であると言える．したがって，今後の課題として得られたスペクトルをさらに分析し有用な情報が得られないかどうか検証していく必要がある．

7.3 対象とする TV 番組映像

本研究では TV ニュース番組を対象として Web との話題構造比較，一般性・受容度の可視化による視聴支援手法の提案を行った．しかしニュース情報は元来十分に洗練されており本提案手法のような品質評価は不必要である可能性がある．今後の課題として，本来対象とすべきバラエティ番組やワイドショーのような TV 番組映像を分析していく必要がある．

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] KeyGraph: “語の共起グラフの分割・統合によるキーワード抽出, 大澤 幸生, ネルス E. ベンソン, 谷 内 田 正 彦, 信 学 会 論 文, Vol” (1999).
[5] Q. Ma and K. Tanaka: “Topic-Structure Based Complementary Information Retrieval for Information Augmentation”, Lecture Notes in Computer Science (APWeb2004), pp. 608–619 (2004).
[6] 甲谷優, 湯本高行, 小山聡, 田島敬史, 田中克己：“Web ページの PageRank 値に基づくローカルコンテンツの品質推定”.
[7] O. Kurland and L. Lee: “PageRank without hyperlinks: structural re-ranking using links induced by language models”, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 306–313 (2005).
[8] G. Salton: “Automatic Information Organization and Retrieval.”, McGraw Hill Text (1968).
[9] A.L., R. Albert：“Emergence of Scaling in Random Networks”, Science, **286**, 5439, p. 509 (1999).