
学術論文

動画のタグ情報を利用した ニュースシーンの選択手法の提案

Selection Technique of News Movies using Tag Information

キーワード：

話題抽出, 動画ニュース, 評価指標, クラスタ, タグ

Keyword：

Topic Extraction, News Movie, Evaluation Index, Cluster, Tag

株式会社NIコンサルティング 中村 浩之

NI Consulting Co.,Ltd Hiroyuki NAKAMURA

電気通信大学 小川 祐樹

University of Electro-Communications Yuki OGAWA

電気通信大学 諏訪 博彦

University of Electro-Communications Hirohiko SUWA

電気通信大学 太田 敏澄

University of Electro-Communications Toshizumi OHTA

要 約

動画ニュースは、人々にとって社会の情報を知るうえでの貴重な情報源である。しかし、日々放送される膨大な動画ニュースのなかから視聴者自身が必要なニュースを取捨選択することは困難となりつつある。本研究では、社会で起きたニュース情報の移り変わりを把握し、視聴者が見たい動画ニュースシーンを選択するための評価指標を構築することを目的とする。我々は、動画ニュースシーンに付随するテキストデータを用いて、同じ内容のシーンを集約したトピックを抽出し、さらに、ニューストピック、及びニュースシーンの特性分析を行うことで、トピックのニュース特性である「流行性」、シーンのニュース特性である「要約性」と「進展性」を指標化する。構築した評価指標の妥当性を検証するために、新聞の記事面積との比較分析や被験者実験を行っている。その結果、提案した指標が、注目されたニューストピックやトピック内の情報をより要約したシーン、既視聴シーンから更新されたシーンを表現できていることを確認している。

2010年7月17日受付 2011年1月23日受理

Abstract

We aim to identify characteristics of news movies that may be various with respect to their topics. The text information of news movies is analyzed to develop similarity graph of news scenes, to identify their topics, and to develop property based indexes. Among the results, the index of the trendy topics, the condensed contents, and the incremented contents among news scenes are developed to identify the property of news movies.

1. はじめに

TV局から放送される動画は非常に膨大である。ドラマや映画、バラエティなどの動画が人々の嗜好を満足するために視聴されるのに対し、動画ニュースは人々が社会の情報を知るための貴重な情報源として視聴される。しかし、多彩なジャンルにまたがり、日々刻々と今までなかった話題が放送され続ける膨大な動画ニュースの中から、視聴者が必要なニュースを効率的に取得することは困難である。我々は、動画ニュースがもつこの問題に着目する。

膨大な動画ニュースのなかから必要な情報を効率的に取得するという問題に対して、TV局によって編集・要約されたニュース番組の視聴や、システムが自動で情報をフィルタリングしてくれる推薦システムの利用などが考えられる。しかし、これらの方法は、情報の効率的な取得や動画ニュースの特性に対する視聴ニーズに対して、十分に応えているとはいえない。

例えば、ニュース番組の視聴においては、番組構成によって、見たくないニュースシーンも含まれる場合がある。この問題については、ニュースを番組単位ではなく、シーン単位で扱うことで、より詳細な選択が可能となる。推薦システムは、他者の行動履歴との類似性（協調フィルタリング）や、利用者の視聴履歴に基づくコンテンツ内容の類似性（コンテンツフィルタリング）により推薦対象となるコンテンツが決定されており（Adomavicius et al. 2005）、視聴者の嗜好と類似したコンテンツを推薦できるという利点がある。ドラマや映画、バラエディ番組など、視聴者がその嗜好に基づいて選択するコンテンツには有効な手法であると考えられる。しかし、協調フィルタリングは、動画ニュースのように日々刻々と入れ替わるコンテンツに対して、コールドスタート問題¹や評価値疎ら問題²を抱えているため利用できない。また、コンテンツフィルタリングは、動画ニュースが社会の情報を知る上での情報源として

視聴されるという特性から、個人の嗜好だけではなく社会的な関心の程度によってニュースの重要度が増減するため有効でないと考える。

視聴者の動画ニュースに対するニーズとして、政治には興味はなくても総選挙の結果を知りたい。あるいは、スポーツに興味はないがオリンピックの結果を知りたいといったニーズが考えられる。また、より短い時間で要約を知りたい、内容が変化したニュースを知りたいなどが考えられる。このように、ニュースシーンを選択するためには、各ニュースシーンがどのような特徴を持っているのか判定する必要がある。しかし、個々の動画ニュースシーンが、視聴者の情報選択という観点においてどんな特性を持っているかは十分に議論されておらず、そのための妥当な評価指標は構築されていない。

本研究の目的は以下のとおりである。すなわち、動画ニュースシーンに付随するタグ情報を用いて、社会で起きたニュース情報の移り変わりを把握し、視聴者が見たい動画ニュースシーンを選択するための評価指標を構築する。具体的には以下の3指標を構築する。

1. 社会の関心の程度を表す指標
2. ニュース内容の網羅性を表す指標
3. ニュース内容の変化の程度を表す指標

我々は、動画ニュースシーンに付随するメタデータを用いて、同じ内容のシーンを集約したトピックを抽出し、さらに、ニューストピック、及びニュースシーンの分析を行うことで、トピックのニュース特性である「流行性（社会の関心の程度）」、シーンのニュース特性である「要約性（ニュース内容の網羅性）」と「進展性（ニュース内容の変化の程度）」の評価指標を構築する。

構築した評価指標の妥当性を確認するため、新聞の記事面積と評価指標との関連、被験者実験と評価指標との関連について分析する。

なお、シーン単位のメタデータについては、動画を各シーンに自動で分割する研究（金寺ら 2005）や、ニュース内容を字幕や音声情報からメ

タ情報として取得する研究（桑野2007）があり、今後このようなシーン単位のメタデータが、必要なニュースを効率的に取得するために利用可能になると考えている。

以降、本稿の構成として、2節では、関連研究として、ニュースの評価指標に関する研究について述べ、本研究の位置づけを示す。3節では、動画ニュースシーン選択について述べる。4節で、動画ニュースシーン選択のための指標を提案し、5節でその評価を行う。6節では、結果に対する考察を行い、7節で結論を述べる。

2. 関連研究

佐藤ら(2005)は、人々の関心の増大によって関連文書数が増加する「情報の広がり」と、そのような注目状態が時間的に継続する「情報の伸び」の観点に着目して話題を抽出する手法を提案している。

菊池ら(2008)は、時系列テキストを対象に、話題の推移を表現したキーワード群を抽出する手法を提案している。電子番組表（EPG）を用いた適用実験により、提案手法が話題の抽出に有効であることを報告している。

張ら(2006)は、トピックの出現頻度が急上昇するところ（バースト時間帯）を抽出し、それと他の時系列データ（小泉首相の支持率、ブッシュ大統領の支持率）とを比較することで、他の時系列データに影響を及ぼすトピックの抽出を行っている。

これらの研究は、付随するテキスト情報の類似度に基づきトピックの抽出を行ったうえで、各トピックが社会の中でどの程度関心を持たれているのかを判定している。佐藤ら（2005）がトピックに含まれる文書の量に基づき話題を抽出しているのに対し、菊池ら（2008）や張ら（2006）は、時事性を考慮し通常よりも出現頻度が多いところを抽出している点で異なる。

しかし、いずれの研究においてもトピック（話

題）の抽出にとどまっており、具体的なシーンの抽出については議論されていない。動画ニュースの視聴を考えたときには、ニュース群からのトピックの抽出だけではなく、内容的・時間的効率性の観点から、各トピックのなかから視聴者のニーズに合ったシーンのみを選択する必要がある。我々は、動画ニュースシーン選択のための指標を構築することを目的とする。

ニュースシーンを考慮した研究として、澤井ら(2008)は、ニュースに付随するテキスト情報の類似性に基づきトピックを抽出したうえで、ニュースシーンの放送順序、ヘッドラインでの紹介の有無、放送時間、放送頻度などの放送状況データを用いて、視聴者にとって重要なニューストピックを抽出している。出現頻度だけでなく、放送順序やヘッドラインでの紹介の有無など、複数のルールを組み合わせているため、組み合わせにより算出コストが増加することが懸念されている。我々は、各ニュースシーンに付随するテキスト情報と放送時間のみを用いて指標を算出する点で、放送順序、ヘッドラインでの紹介の有無などを考慮する澤井ら（2008）の研究と異なる。

3. 動画ニュースシーン選択

本節では、視聴者の動画ニュースシーン選択におけるトピック選択とシーン選択について述べる。

3.1 トピック選択とシーン選択

我々は、視聴者が必要なニュースシーンを効率的に選択するために、動画ニュースにおけるトピック（話題）の特徴付けと、トピック内でのシーンの特徴付けが必要であると考えます。

TV局から放送される動画ニュースは非常に膨大であるが、そのなかには同じ内容のニュースが多く含まれている。動画ニュースの視聴を考えたとき、一般的に視聴者は複数のトピックのニュースを効率的に視聴したいと考えられる。そのため、まずは同トピックのニュースであることを判断す

るためにも、内容に基づいたトピックの分類が必要であると考えられる。

しかし、一日に放送される動画ニュースにおける話題は多種多様に存在するため、視聴者が必要なトピックを選定するためには、トピックの分類だけでなく、複数のトピックのなかからどのトピックを選定するかといった選択基準が必要になる。

さらに、トピックを選択しても、同トピック内に含まれるニュースシーンであっても、報道する側の視点や時間変化などによって、各ニュースシーンの放送時間や内容は異なる。そのため、トピック内においても、そのシーンの特徴付ける必要があると考えられる。図-1は、このような視聴者における動画ニュース選択の例を示したものである。はじめに複数のトピックの中からTopicBを選択し、次にTopicBに含まれるニュースシーンからSceneEを選択している。

以上をふまえ、我々は、視聴者が動画ニュースを効率的に選択できるようにするために、動画ニュースにおけるトピックの特徴付け、またトピック内でのシーンの特徴付けを行う。

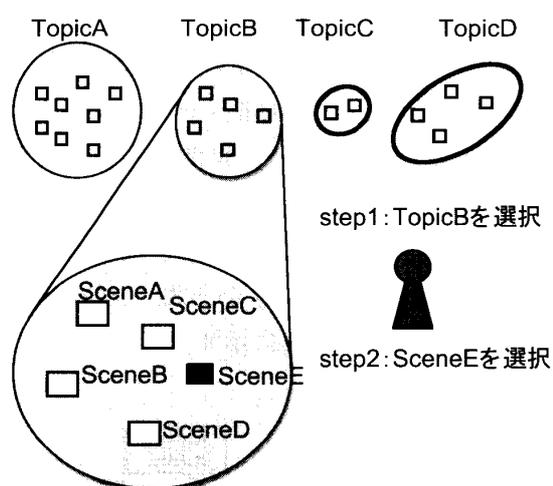


図-1：トピック選択とシーン選択の例

3.2 トピック選択の基準

視聴者によるドラマやバラエティなどの番組選

択は、視聴者の番組内容に対する嗜好によって選択が行われると考えられる。しかし、ニュースコンテンツを対象としたときには、そのニュースの社会的な注目度といった観点から、重要な判断基準になると考えられる。例えば、ニュースとして報道される情報は、自分の知的欲求を満たすためだけでなく、他者とのコミュニケーションを円滑にするためにも用いられる。このような用途を目的とした場合、より多くの人に注目されている流行性の高いニュースの選択が必要になる。

視聴者がこのようなニュースのトピックを選択できることを支援するためには、トピックに対する社会的な注目度といった観点からのニューストピックの特徴付けが有用であると考えられる。日々放送されるニュースは、社会的に注目されているニュース(報道数が多いニュース)とそうでないニュース(報道数が少ないニュース)がある。報道数によりトピックの評価を行うことで、流行性を示すことが可能となると考える。

3.3 シーン選択の基準

視聴者がニュースを視聴する理由の一つとして、情報収集といった目的が考えられる。例えば、ある話題に関して、詳細な情報までは必要ないが、概要だけ簡単に知っておきたいなどの視聴理由が考えられる。この場合には、より多くの情報を短時間に取得するために、放送時間が短く、かつ内容がまとめられている要約性の高いニュースシーン選択が必要となる。

また、視聴者の望むニュースシーンは要約性の高いニュースだけでなく、興味のあるトピックのなかでも進展のあったニュースを知りたいといった視聴理由も考えられる。この場合、新たな内容情報(差分)を比較したうえで、新たな情報が付加されたニュースシーン選択が必要になる。

これらをふまえると、視聴者がトピックのなかから必要なニュースシーンを選択できることを支援するためには、ニュースシーンの要約性、あるいは進展性といった観点での特徴付けが有用であ

ると考える。

4. 提案手法

3節において、動画ニュースを選択するためにトピックを抽出する必要があること、トピック及びシーンを選択するために指標が必要であることを示した。本節では、トピック及び各指標の具体的な抽出方法について提案する。また、提案手法をメタデータに適応した場合の結果を例示する。なお、提案手法の評価については5節で述べる

なお、使用したメタデータ³は、2009/03/23~03/29と2009/12/17~12/23の二つの期間内にTVキー局で放送されたニュースシーンに付随するテキスト情報（番組名、ヘッドライン、Memo）と放送時間である。Memoは、実際に番組を見て人手でその内容を記述したものである。将来的には、音声認識技術や画像認識技術により自動化可能な作業と考える。

4.1 トピック抽出

本研究では、トピックの抽出をNewman法(Newman 2004)によるクラスタリングによって行う。Newman法によるトピック抽出はWeblog上のニュース記事を用いたトピック抽出手法としても使われおり(内田 2006, 松永 2007), クラスタ数が自動で決定できるという利点を持っている。具体的な抽出に関しては、まず、放送された各ニュースシーンのテキスト情報に対して形態素解析を行い、テキスト情報に含まれる名詞をタグとして抽出する。そして、抽出したタグのtf-idf値⁴を各シーンの特徴量とし、シーン同士のcos類似度⁵を求める。このcos類似度が閾値0.4以上のものを用いて、類似するシーン同士が結びつくシーン類似度ネットワークを作成する。最後に、シーン類似度ネットワークのクラスタリングをNewman法により行い、各クラスタをトピックとして抽出する。トピックの名前付けは、tf-idf値の高い用語を用いて行っている。

図-2は、2009/12/18におけるシーン類似度ネットワークと抽出したトピック(点線で囲ったシーン集合)を示している。

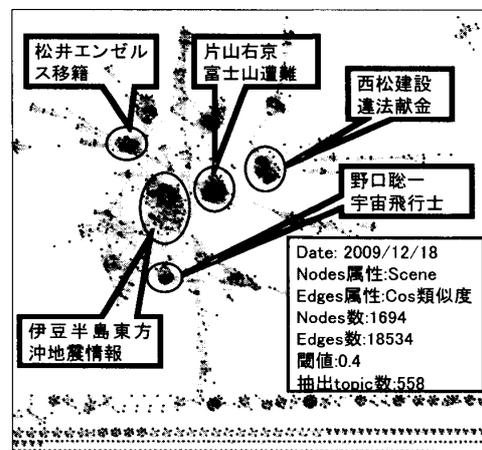


図-2 トピック抽出の例 (2009/12/18)

4.2 流行性指標 / trend

トピック*i*に含まれるシーン数の多寡により、流行性trendを表現する。

$$\text{trend}(i) = \frac{\text{トピック}i\text{のシーン数}}{\text{全トピックに含まれる平均シーン数}}$$

1日ごとのトピック数やシーン数が異なるため、1トピックあたりの平均シーン数で正規化を行う。トピック内に含まれるシーン数が少なければ、trend値が小さくなり、世間の注目度が低い事を意味する。逆に含まれるシーン数が多ければ、trend値が大きくなり、世間の注目度が高い事を意味する。trend値が大きかったトピック群のtrend値推移を図-3、図-4に示す。

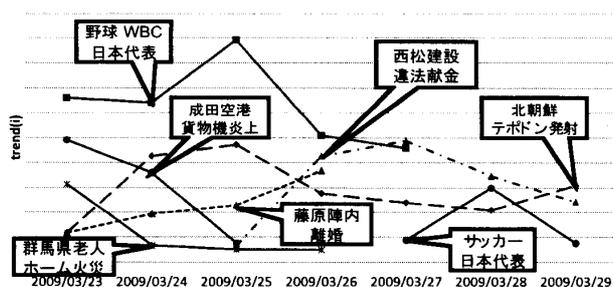


図-3: 2009/3/23~3/29におけるtrend値の推移

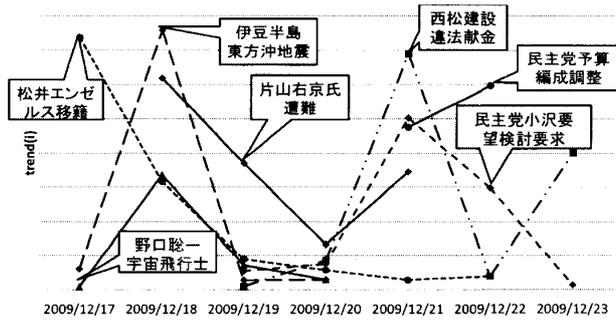


図-4:2009/12/17~12/23におけるtrend値の推移

4.3 要約性指標 / abst

トピックiの重要語と、トピックi内のシーンjのタグの重なり(の程度及び放送時間)を用いて要約性(abst)を表現する。

$$abst(j) = TAG_{ij} \text{の} tfidf \text{値の合計} \cdot e^{-aTime_j}$$

TAG_{ij}は、トピックiの重要語でありかつシーンjにも含まれるタグ集合である(図-5)。

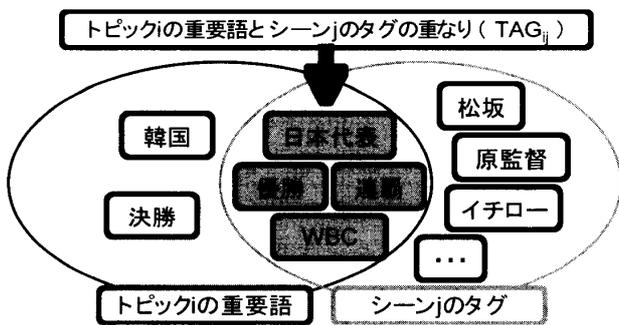


図-5: TAG_{ij}の例

トピックiの重要語を抽出するために、TFIDF_mを用いる。

$$TFIDF_m = \frac{\text{トピックiに含まれる単語mの} tfidf \text{値の合計}}{\text{トピックiに含まれるシーン数}}$$

TFIDF_mは、トピックiに含まれるタグmのtf-idf値を合計し、トピックiに含まれるシーン数で割ったものである。トピックiに含まれるすべてのタグに対してこの値を算出し、上位10個をトピックiの重要単語とする。トピックの重要語をよ

り多く含むシーンは、よりそのトピックの内容を網羅していると考えられる。しかし、動画ニュースシーンは、それぞれ放送時間が異なり、放送時間の長い報道は情報量が増えることが推測される。そのため、放送時間が長いシーンほど網羅性が高くなる可能性が高い。そこで、e^{-aTime}を用いて補正を行う。Time_jは、シーンjの放送時間(秒)であり、aは調整係数である。今回は、視聴者が1日にニュースを見る時間を20~30分程度、視聴するトピックを5~10個と仮定し、a=0.02と設定している。重要語のtf-idf値を合計し、放送時間により補正することで、シーンの要約性を表現する。

表-1: 要約性指標の算出例

ヘッドライン	Memo	Abst
【野球】WBC日本代表・優勝	日本代表が優勝し、2連覇を達成した。WBC日本代表・監督・原辰徳氏、イチロー、岩隈久志、松坂大輔が優勝の感激を語った。MVPは松坂大輔。ハイライトシーンを紹介。(中継)米国ロサンゼルス・ドジャーススタジアム。【出演】恵俊彰、久保田智子、榎原寛己、佐々木主浩【解説・コメンテーター】山田五郎氏、山口もえ【コメント】原辰徳氏(WBC日本代表・監督)、イチロー、岩隈久志、松坂大輔	1.807
<スポーツHEADLINE>【野球】WBC日本代表・決勝進出	準決勝・日本代表-米国。【会見】原辰徳氏(WBC日本代表・監督)	0.894
<ズーム★スポーツ>【野球】WBC決勝・日本代表-韓国	日本代表選手が会場入りした。(中継)韓国ソウル、米国ロサンゼルス【出演】熊谷章洋、青池奈津子	0.115
【最新情報】		

要約性指標の算出例を表-1に示す。該当トピックの主な内容は、『WBC日本代表が、決勝で韓国に勝利し、優勝2連覇を達成した。』である。この文章を構成するタグの出現頻度が大きいシーンほど、abst値は高く評価されている。

4.4 進展性指標 / progre

前日のトピックiと当日のトピックiの重要語を比較し、新たに重要語となったタグと当日のi内のシーンjの重要語の重なり(の程度)を用いて進展

性progreを表現する。

$$progre(j) = TAG_{ij} \text{のtfidf値の合計}$$

TAG_{ij} は、当日のトピックiの重要語でありかつ前日より重要度 ($TFIDF_m$) が上昇し、さらにシーンjに含まれるタグ集合である(図-6)。

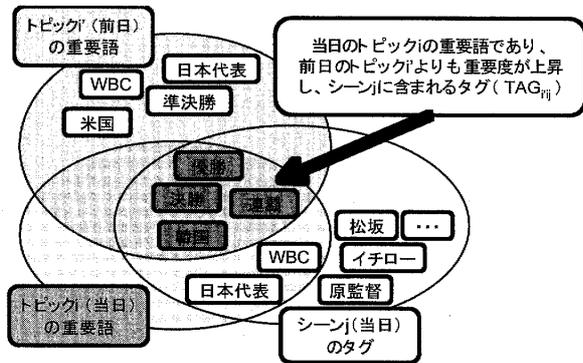


図-6: TAG_{ij}の例

当日のトピックiのTFIDF_m値上位10個のうち前日のトピックiのTFIDF_m値より上昇したタグを抽出する。そのようなタグをより多く含むシーンは、新たな重要な情報をより多く含んでいると考える。新たに発生した重要語のtf-idf値を合計することで、シーンの進展性を表現する。

進展性指標の算出例を表-2に示す。前日の重要語は、「WBC, 日本, 準決勝, 勝利, 米国」などである。当日の重要語は、「WBC, 日本, 勝利, 優勝, 連覇韓国」などであり、前日になかった「優勝, 連覇, 韓国」などの用語が含まれるシーンのprogre値が高い値となっている。

ヘッドライン	Memo	Progre
【野球】WBC 日本代表・優勝	決勝・日本代表-韓国。日本代表が勝利し、2連覇を達成した。【コメント】原辰徳氏(WBC日本代表・監督), イチロー【会見】松坂大輔, 岩隈久志, ダルビッシュ有, 内川聖	7.383
【野球】WBC 日本代表・優勝	号外の一面を紹介	1.251
【野球】WBC 日本代表・決勝進出	【会見】原辰徳氏(日本代表・監督)	0

表-2:進展性指標の算出例

5. 提案手法の評価実験

本節では、4節で提案した各指標が妥当であるか評価実験を行う。

5.1 流行性指標/trend

流行性指標の妥当性を検証するために、trend値と新聞の記事面積との比較を行う。木村ら(2002)は、複数の新聞の記事面積の分析から「世論」とよばれる「現実」の形成や振る舞いを見る事ができると述べている。我々は、記事面積が広いニュース程、世の中に注目されていると考える。本研究では、朝日新聞と読売新聞の2つの新聞記事面積を用いて、trend値の高いトピックが世の中の注目されているニューストピックであるかを検証する。

分析対象のトピックは、図-3中の「WBC」「成田空港貨物機炎上」「群馬老人ホーム火災」「西松違法献金」「北朝鮮テポドン発射」の5つのトピック(1週間分)である。対象トピックの記事面積を集計し、trend値の推移との相関を分析した。トピック記事面積とtrend値の散布図を図-7に示す。相関係数は0.756であり高い相関を示した。これにより、流行性指標が社会で注目されているニュースであることを示す指標であることが確認できる。

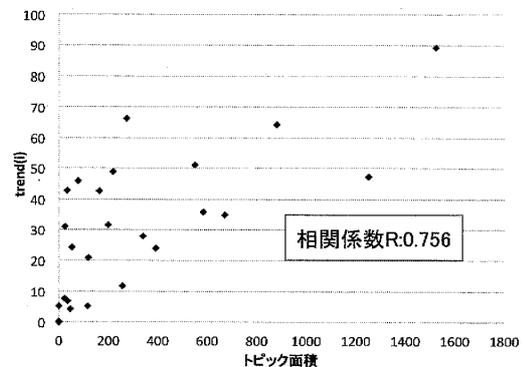


図-7: トピック面積とtrend値の関係

5.2 要約性指標/abst

要約性指標の妥当性を示すために、被験者実験

を行う。被験者（10人）に同一トピックのabst値の異なるニュースシーンを視聴してもらい、要約性の程度を判断してもらう。その判定結果とabst値の大きさを比較する。視聴するトピックは、『民主党予算編成調整』（政治・国際面）、『野口聡一宇宙飛行士』（サイエンス面）、『サッカー』（スポーツ面）、『片山右京氏遭難』（社会面）の4トピックであり、それぞれ3シーンずつ視聴する。

被験者による判定結果とabst値の順位相関検定を行った結果を表-3に示す。検定結果より、民主党予算編成とサッカー、片山氏遭難に関するトピックに有意な相関が確認されている。よって、要約性指標はある程度妥当であると考えられる。

表-3：要約性指標の順位相関検定結果

対象トピック	民主党 予算編成	野口聡一 宇宙飛行士	サッカー	片山氏 遭難
Kendall's τ	0.373	0.160	0.653	0.450
検定結果	$P < 0.01$	$P > 0.05$	$P < 0.01$	$P < 0.01$

5.3 進展性指標/progre

進展性指標の妥当性を検証するために、要約性と同様に被験者実験を行っている。被験者（10人）に前日のニュースを視聴してもらった後、当日のニュースを視聴してもらい、進展性の程度を判断してもらう。その判定結果とprogre値の大きさを比較する。視聴トピックは、要約性指標に用いたトピックを用いている。

被験者による判定結果とprogre値の順位相関検定を行った結果を表-4に示す。検定結果より、民主党予算編成と野口聡一宇宙飛行士、サッカーに関するトピックに有意な相関が確認されている。よって、進展性指標はある程度妥当であると考えられる。

表-4：進展性指標の順位相関検定結果

対象トピック	民主党 予算編成	野口聡一 宇宙飛行士	サッカー	片山氏 遭難
Kendall's τ	0.133	0.357	0.357	0.010
検定結果	$P < 0.01$	$P < 0.01$	$P < 0.01$	$P > 0.05$

6. 考察

本節では、提案指標の特性及び利用可能性について考察する。

6.1 トピック選択に関する指標

新聞記事との比較により、流行性指標が妥当であることが示されている。trend値の推移傾向としては、急に上昇・下降するタイプと、徐々に変化するトピックがあり、基本的には前者が多く見られる。例えば、成田空港の貨物機炎上や北朝鮮情勢に関するニュースなどは、その事件が起こった日にtrend値が急上昇し最も高い値を示している(図-3)。このような偶発的なニュースは、ニュースの事件性により一時的に流行性が高くなるものとする。

一方、WBCや西松献金事件に関するニュースなどは、流行性が徐々に変化している(図-3)。WBCに関するトピックの場合、23日に準決勝で米国に勝利し、24日に韓国との決勝戦に勝利している。25日に、2連覇を祝福するニュースが多く報道され、26日に凱旋帰国を果たしており、一連の流れの中で「WBCで2度目の優勝をした」という段階が最も高い流行性を示している。このような一連の流れがある程度推測できるニュースは、トピックに対する社会の関心度と連動して、流行性が徐々に変化するものとする。

なお、本研究では、トピックの注目の程度を、佐藤ら（2005）と同様にトピックに含まれるニュースシーンの量に基づき算出している。1日に報道されるニュースの量はある程度決まっているため、菊池ら（2008）や張ら（2006）のようにアルゴリズムを用いて時事性を考慮しなくても、時事性を十分に表現できていると考える。

6.2 シーン選択に関する指標

5.2, 5.3節の被験者実験により、要約性指標と進展性指標が妥当な指標であることが示されている。ただし、4トピック中、1トピックに関して

は有意な差は確認されなかった(表-3、表-4)。要約性指標において、野口聡一宇宙飛行士に関するトピックの順位相関検定が有意にならなかった理由は、提案手法が「ソユーズ打ち上げ」に関連する用語を含む比較的放送時間の長いシーン(2分23秒)を要約性が高いと判定したのに対し、複数の被験者が放送時間の短いシーン(24秒)を要約性が高いと判断したためである。提案手法において、視聴者が1日にニュースを見る時間を20~30分程度、視聴するトピックを5~10個だと仮定し、時間の補正($a=0.02$)を行っているが、視聴者によって時間に対する重要度は異なる。視聴者ごとに a の値を決定することで、個々の視聴者にとって適切な長さのニュースの発見が可能になると考える。 a の決定方法については、時間に対する重要度を視聴者本人に設定してもらう方法や、視聴履歴に基づいてシステムが設定する方法などが考えられる。今後の検討課題である。

また、進展性指標において、片山氏遭難に関するトピックの順位相関検定が有意にならなかった理由は、トピックの発生(事故の発生)から4日経過し、「遭難に関する専門家のコメント」、「関係者のインタビューの様子」、「当面活動中止に関する報道」など様々な角度での報道がなされており、被験者により進展したと感じるシーンが異なっていたためと考える。このように、そもそも被験者によって進展したと感じるシーンが異なる場合は、本手法では取り扱えないと考える。一方、同様にトピックの発生から4日経過している民主党予算成立に関するトピックについては、記者クラブを通じた情報を用いて放送されるトピックであったため、指標が有効に機能したものと考える。

6.3 推薦システムへの提案指標の適用

従来の推薦システムは、他者の行動履歴との類似性(協調フィルタリング)や、利用者の視聴履歴に基づくコンテンツ内容の類似性(コンテンツフィルタリング)により推薦している。しかし、動画ニュースは、常に新しいシーンが発生するた

め、協調フィルタリングのコールドスタート問題や評価値疎ら問題を解決できない。本手法であれば、タグデータが付与された時点で指標を算出可能であり、この問題を解決できると考える。

また、動画ニュースを選択する際には、ニュース内容に対する好みといった理由だけでなく、社会的に注目されているニュースを知っておきたいといった視聴理由や、ニュースの概要を内容的・時間的に効率的に取得したいといった視聴理由が存在する。コンテンツフィルタリングは、過去の視聴履歴に基づくコンテンツ内容の類似性を用いて推薦するため、このような視聴理由には対応できない。

本研究で提案した流行性指標は、注目度の高いニュースシーンの程度、また要約性指標はトピック内での内容と時間の効率性、進展性指標はシーンに対して新たな重要情報が付加された程度を表わしており、ニュースというコンテンツに対する視聴ニーズに対応した指標化である。このため、従来の内容嗜好をベースとした推薦方式にみられる適合フィードバックによる推薦方法の枠組みに対して、本指標を取り入れることが可能となり、動画ニュースというコンテンツに対するニーズを考慮した推薦システムの構築に応用できると考えられる。

ただし、本研究で提案した指標は、コンテンツフィルタリングを完全に否定するものではない。例えば、オリンピックの開催時期でも、オリンピックに興味がない視聴者は存在する。そのような視聴者に対しては、流行性に関わらず嗜好性に基づいてトピックを推薦することも有効と考える。また、特定のトピックに関するニュースだけは視聴したいと考える視聴者も存在する。そのような視聴者には、嗜好性と流行性の両方を考慮したハイブリッドの推薦手法が有効と考える。各視聴者にあわせて、どの指標に重みをもたせるかの検討が今後の課題である。

7. 結論

我々は、動画ニュースシーンに付随するメタデータを用いて、動画ニュース選択のための新たな指標を構築した。ニュースを番組単位ではなく、シーン単位で扱うことで、より詳細な選択が可能となった。トピック選択を目的とする指標として、社会の関心の程度を表す流行性指標を構築した。シーン選択を目的とする指標として、ニュース内容の網羅性を表す要約性指標、変化の程度を表す進展性指標を構築した。指標の妥当性を新聞紙面との比較および被験者実験によって示した。提案指標により、視聴者は大量の動画ニュースシーンから、注目されたニュースや、トピック内の情報をより要約したニュース、既知情報からの差分となるニュースを選択できるようになる。

今後は、3つの指標を組み合わせ、3次元空間に記事をマッピングして分類する手法や、これらの指標を用いた動画ニュース推薦システムの検討が課題と考える。

本研究では、手動でタグ付けされているデータを用いて分析を実施している。より広範な利用のためには、自動タグ付けの技術が必要となる。今後の技術開発に期待したい。また、trend, abst, progre の定義はいずれもアドホックであり、他の指標を検討する余地がないわけではない。今後よりよい指標の検討をさらに続けたい。

謝 辞

本研究を行うにあたり、株式会社ミディー様よりデータ提供だけでなく貴重な御助言と、実験を行う上で並々ならぬご支援、ご協力を頂きました。ここに記し感謝いたします。

注

- ¹ コールドスタート問題とは、データ収集初期にデータ量が少なく、適切な推薦ができない問題である。
- ² 評価値疎ら問題とは、注目されないコンテンツには評価値が付与されないため、適切な推薦ができない問題

である。

- ³ 今回使用したメタデータは、専門業者より購入したものである。よって、コストの問題を解決できれば誰もが利用できるデータである。
- ⁴ tf-idf値は、単語の重要度を表し、以下の通り求める。

$$tfidf=tf \cdot idf$$

$$tf_i = \frac{n_i}{\sum_k n_k}$$

$$idf = \log \frac{|D|}{|\{d:d \ni t\}|}$$

ni:文書n中の単語iの出現頻度

|D|:総ドキュメント数

|\{d:d \ni t\}|:単語を含むドキュメント数

- ⁵ cos類似度 ($\text{sim}(s_a, s_b)$) は、シーン同士の類似性の程度を表し、以下の通り求める。

$$\text{sim}(s_a, s_b) = \frac{\sum S_{i,a} S_{i,b}}{\sqrt{\sum (S_{i,a})^2} \sqrt{\sum (S_{i,b})^2}}$$

s_x : シーンx

$S_{x,y}$: シーンxにおけるタグyの特徴量

参考文献

- Adomavicius, G. and Tuzhilin, A. (2005), Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.6, pp.734-749.
- 張一萌, 何書勉, 小山聡, 田島敬史(2006), 時系列データに意味的に関連するニューストピックの発見, 日本データベース学会論文誌, Vol. 5, pp.1-4.
- 金寺登, 隅田飛鳥, 池端孝夫, 船田哲男(2005), ビデオ教材作成支援を目的とした講義音声によるシーン分割, 電子情報通信学会論文誌D-I, 情報・システム, I-情報処理 J88-D-I(5), pp.977-984.
- 菊池匡晃, 岡本昌之, 山崎智弘 (2008), 階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出, 日本データベース学会論文誌 Vol.7, No.1, pp.85-90.
- 木村洋二, ハンナロンチャーン, 板村英典 (2006),

- 日本の4大新聞における皇室報道の比較研究：皇太子さまの「人格否定」発言を事例として，関西大学社会学部紀要，Vol.37, No.3, pp55-106.
- 桑野秀豪，松尾義博，川添雄彦 (2007)，映像・音声認識，自然言語処理の適用によるメタデータ生成の作業コスト削減効果に関する考察，映像情報メディア学会誌:映像情報メディア，Vol.61, No.6, pp.842-852.
- 松永拓 (2007)，キーワードの出現に基づくブログコミュニティ抽出とオピニオンリーダーの発見，電子情報通信学会第18回データ工学ワークショップ論文集 (DEWS2007)，C3-7.
- Newman, M.E.J. (2004), Fast algorithm for detecting community structure in networks, Phys. Rev. E, Vol.69, 066133.
- 佐藤吉秀,川島晴美, 佐々木努, 大久保 雅且(2005), 文書の類似度と新鮮度に基づく話題語抽出, 情報処理学会研究報告. 自然言語処理研究会報告, No.1, pp.29-35.
- 澤井里枝, 妹尾宏, 鹿喰善明(2008), ニュースダイジェスト作成システムにおける重要度算出手法の評価, 情報処理学会研究報告. データベース・システム研究会報告, No.7, pp.155-162.
- 内田誠, 柴田尚樹 (2006), ブログ記事ネットワークからの emerging topic の抽出と可視化, 人工知能学会全国大会, 3D2-3.