Wikipedia を利用したニュース映像アーカイブへの自動索引付け

奥岡 知樹† 高橋 友和†† 出口 大輔† 井手 一郎†,††† 村瀬 洋†

† 名古屋大学大学院情報科学研究科 〒464-8601 愛知県名古屋市千種区不老町
†† 岐阜聖徳学園大学経済情報学部 〒500-8288 岐阜県岐阜市中鶉 1-38
††† 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: † {okuoka,ide,murase}@murase.m.is.nagoya-u.ac.jp, †† ttakahashi@gifu.shotoku.ac.jp, ††† ide@nii.ac.jp

あらまし テレビやインターネット上での映像コンテンツの増加により，映像アーカイブを効率良く閲覧・検索する技術が求められている．そこで本稿では，資料映像として利用価値の高いニュース映像に注目し，Wikipedia を利用したニュース映像アーカイブの閲覧支援技術を提案する．テキスト情報の類似度評価を基に，各映像に対して Wikipedia エントリによる自動索引付けを行う．索引付け結果を利用し，各エントリに関連する映像群を抽出，提示する．実験により適合率 86%，再現率 79%で索引付けが行えることを確認した．また，話題の変遷の理解や放送日が離れた映像間の関連発見など, クローズドキャプションのみでは得られない情報を得られることを確認した．

キーワード Wikipedia，ニュース映像，映像アーカイブ，自動索引付け

Automatic Indexing of a News Video Archive with Wikipedia Entries

Tomoki OKUOKA† Tomokazu TAKAHASHI†† Daisuke DEGUCHI†

Ichiro IDE†, ††† and Hiroshi MURASE†

† Nagoya University Graduate School of Information Science Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601 Japan
†† Gifu Shotoku Gakuen University Faculty of Economics and Information 1-38, Nakauzura, Gifu-shi, Gifu, 500-8288 Japan
††† National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
E-mail: † {okuoka,ide,murase}@murase.m.is.nagoya-u.ac.jp, †† ttakahashi@gifu.shotoku.ac.jp, ††† ide@nii.ac.jp

Abstract Following the increase of video contents on TV or the Internet, efficient techniques to browse and search video data in an archive are needed. We propose a method to support browsing a news video archive with the help of Wikipedia because news videos are important as video contents. First, videos are automatically indexed by Wikipedia entries by means of evaluating the similarity of text information. Using those indices, we extract and present video groups via an interface, that correspond to each Wikipedia entry. Through experiments, news videos were accurately indexed by Wikipedia entries with a precision of 86% and a recall of 79%. In addition, we confirmed that we could obtain information which could not be obtained by closed captions, such as understanding the transition of topics or discovering the associations between videos which are broadcasted on distant days.

Keyword Wikipedia, News Video, Video Archive, Automatic Indexing

1. はじめに

近年，YouTube(*1)などの動画共有サイトの普及により，誰でも手軽に映像アーカイブを利用することが可能となった．また NHK アーカイブス(*2)など，過去に放送された映像を再利用するという取り組みも行われており，このような映像アーカイブを効率的に閲覧・検索する技術が求められている．その中でもニュース映像は資料映像として利用価値が高い.そこで我々は，ニュース映像アーカイブの効率的な閲覧技術に注目している．

ニュース映像の構造解析や閲覧技術に関しては，多くの研究がなされている．その多くは映像に付与された文字放送字幕テキスト（Closed Caption；以下 CC）を利用し，それらのテキスト情報の類似度を基に映像間の関連を分析している．井手らは，強く関連したニュース映像を時系列に連鎖することによりトピックスレッド構造を構築し，映像アーカイブの構造解析や閲

＊1：http://www.youtube.com/
＊2：http://www.nhk.or.jp/archives/

覧インタフェースを実現した[1]．しかし各ニュース映像が表す内容が明示されておらず，あるニュースイベントに関連する映像を閲覧したい場合には適切ではない．またニュースの中には，長い期間を通して少しずつ放送されるトピックも存在し，従来の技術ではこのような場合への対応が困難である．

そこで本稿では，オンライン百科事典として有名なWikipedia(*3)に注目し，Wikipediaを利用したニュース映像アーカイブの閲覧支援技術を提案する．まずテキスト情報の類似度評価を基に，各映像に対してWikipediaエントリにより自動索引付けを行う．次にそれを利用して各エントリに関連する映像群を抽出，提示する．この際，文献[1]のトピックスレッド構造中の各映像に対して索引付けを行うことで，あるニュースイベントに関連した映像群を時系列に閲覧することが可能となる．また従来は計算量的に処理が困難であった，放送日が離れた映像間の関連を発見することも可能となる．

索引付けの際にWikipediaを利用することの利点は，以下に示す2つの特長による．1つ目はWikipediaエントリに関して，表記と概念が一対一に対応していることである．CCにおいては「自由民主党」，「自民党」と様々な表記で使用される概念であっても，Wikipediaでは「自由民主党(日本)」というエントリで扱われる．これにより索引の表記のゆれを解消し，重複のない索引付け・閲覧が実現できる．そして2つ目に，コンテンツの網羅性がある．一般にニュースイベントに関して発端から終息までの一連の流れが詳細に説明されており，閲覧技術の構築に有用であると考える．

以降，第2章で関連研究を紹介した後，第3章でWikipediaを利用したニュース映像アーカイブへの自動索引付けに関する処理の詳細を述べる．続く第4章で，対応付けの精度に関する実験や抽出結果の例を紹介し考察を述べる．最後に第5章において今後の課題を検討し，本稿をまとめる．なお，以降“ストーリ”とは，一つのイベントを扱った，ニュース映像の意味的な最小単位を表わす[2]．

2. 関連研究

2.1. ニュース映像の閲覧に関する研究

ニュース映像の閲覧支援を目的とした研究として，ニュース映像の時系列意味構造解析に関する研究が多くなされている．最も単純な手法として，特定のトピックに関連するストーリを時系列に直線状に連ねる方法が考えられる[2][3]．しかし直線構造では同一トピックであっても同時並行して進む個別の話題の流れを表現できない．これに対しWuらは，特定のトピックに関連するストーリを集めたクラスタにおいて，時系列の前後関係と話題の変化に応じて2分グラフを構築する手法を提案した[4]．しかしこの方法でも，新規ストーリ同士の関係は時系列の前後関係のみであり，同時並行して進む個別の話題の流れを表現できない問題があった．そこで井手らは，同時並行して進む話題の流れを表現する時系列意味構造：トピックスレッド構造を抽出する手法を提案した[1]．トピックスレッド構造の例とそれを利用した閲覧インタフェースを図1に示す．ここでトピックスレッド構造中の各ノードは，ストーリ分割後の各ニュース映像を表す．また井手らはこの研究の中で，構造中の局所的な意味的まとまりの抽出も行った．しかし，抽出されたまとまりがどのようなニュースを表すのかを明示することは行われなかった．またトピックスレッド構造を構築する際，放送日の近い映像を中心として関連付けを行うため，長い期間を通して少しずつ放送されるようなニュースには対応できなかった．そこで我々は，Wikipediaという外部の情報資源を利用することで，これらの問題の解決及び新たなニュース映像アーカイブの閲覧支援技術を目指す．

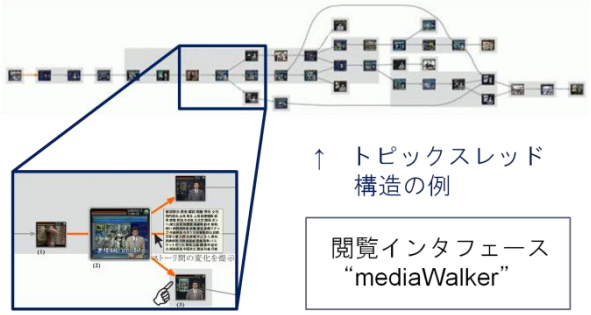

図1　トピックスレッド構造に関する研究

その他，時系列意味構造解析にとらわれずニュース映像の可視化を目指した研究も多くなされている．その代表的なものとして，Rautiainenらによるcluster-temporal browsing[5]や，Snoekらによる閲覧インタフェースMediaMill[6]などが挙げられる．

2.2. Wikipediaの利用に関する研究

Wikipediaは知識抽出のための有用なコーパスとして，人工知能を始めとした様々な分野で研究，応用されつつある．その中でも，Wikipediaの持つ豊富なコンテンツ量を生かした，連想シソーラス辞書の自動構築技術に関する研究が多くなされている．中山らはWikipediaに対してWebマイニングの手法を適用することでシソーラス辞書の自動構築を行った[7]．

＊3：http://ja.wikipedia.org/wiki/Wikipedia/

シソーラス辞書の構築以外にも，Wikipedia を利用した研究は増えつつある．例えば，Wikipedia の情報を様々なマルチメディア情報と対応付けることで，より高度な処理や情報発見を目指した研究などである．川場らは Wikipedia エントリをブログサイトと対応付け，Wikipedia カテゴリ空間におけるブログサイトの分布推定を行った[9]．今後も Wikipedia のような，多くのユーザの共同作業により構築された情報（folksonomy）を利用する研究が注目されるのではないかと考える．

3. Wikipedia を利用した自動索引付け

3.1. 処理の流れ

処理の流れを図 2 に示す．ニュース映像に関しては，付随する CC からトピックスレッド構造を構築する．この際に用いる手法は文献[1]と同様である．一方 Wikipedia に関しては，ニュースに関連するエントリ（以下，ニュース関連エントリ）を抽出する．そして両方の出力結果を利用し，CC と Wikipedia エントリとの間の類似度評価により対応付けを行う．この際，各ニュース関連エントリのテキストから日付情報を抽出して類似度評価の対象を限定した後，トピックスレッド構造を利用し索引の補完を行う．これにより対応付け精度の向上を図る．最後に各エントリに関連する映像群を抽出する．

3.2. ニュース関連エントリの抽出

Wikipedia の全エントリを利用して索引付けを行うのは困難である．その最大の理由は，CC との対応付け精度の低下を招くためである．また 2008 年 11 月 27 日時点でのエントリ数は約 100 万件であり，今後も増加すると考えられ，全てを処理するための計算量も問題である．そこでニュースに関連するエントリのみを抽出する．ニュース関連エントリの抽出の際に，以下に用いる 2 つの特徴を利用した．1 つ目は，ニュース関連エントリのテキスト中に Wikipedia の姉妹プロジェクトであるウィキニュース(*4)へのリンクが存在する場合が多いことである．また 2 つ目は，ニュースの真実性を証明するために，一般のニュースサイト中の記事の URL を参考文献として引用する場合が多いことである．以上の特徴を利用して，Wikipedia からニュース関連エントリを抽出する．

3.3. CC と Wikipedia エントリの対応付け

テキスト情報の類似度評価を行い，CC と Wikipedia エントリを対応付ける．まず CC と Wikipedia エントリのテキストを形態素解析し，名詞の出現頻度ベクトルを作成する．そして両者のコサイン類似度を算出し，

＊4：http://ja.wikinews.org/wiki/メインページ/

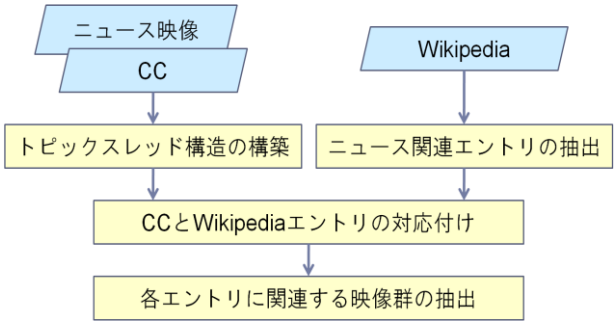

図 2　処理の流れ

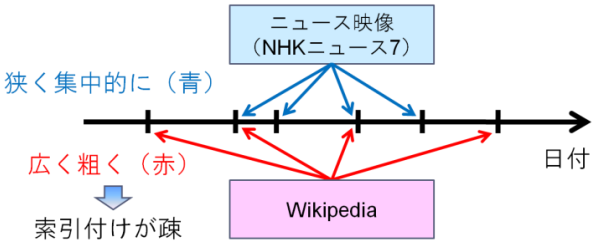

図 3　ニュースの取り上げ方の違い

しきい値を超えればそれらを対応付ける．本稿においては Wikipedia エントリのテキスト全体を使用してベクトルを作成したが，節ごとに作成することもできる．しかしこの場合，一つの節中の名詞が極端に少なくなり，対応付け精度が低下することが多い．この問題を解決するために検討が必要である．

抽出された全てのニュース関連エントリと全ての CC との類似度評価を行った場合，対応付け精度の低下を招く．そこで以下の方法で対応付けを行い，精度向上を目指す．

3.3.1. 日付情報の抽出

ニュースにおいて事象の生起日「いつ（When）」の情報は重要である．そこで Wikipedia エントリのテキスト中から日付情報（****年**月**日）を抽出し，それにより類似度評価の対象期間を限定することで対応付け精度の向上を図る．文中で日付に関する情報が出現する場合，年，月などの情報が省略されることが多い．例えば「2008 年 8 月 25 日から 28 日にかけて…」のような場合である．本研究では，直前に出現した年，月の情報を利用し，このような省略を補完する．先程挙げた例では，「2008 年 8 月 25 日」，「2008 年 8 月 28 日」という日付情報が抽出される．

しかしこの手法には問題が存在する．それはニュース映像と Wikipedia とで，ニュースの取り上げ方が異なる場合である．その概念図を図 3 に示す．ニュース映像はあるトピックに対して，注目されている時期に集中的に取り上げることが多い．それに対して Wikipedia は，一般にニュースイベントの発端から終息

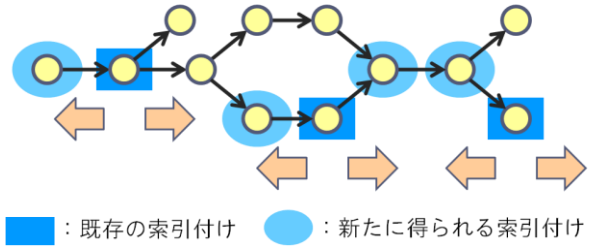

図 4　索引の補完

までを日付情報を含めて網羅的に説明するが，注目されている時期の日付情報を集中して記述することは少ない．そのため日付情報により類似度評価の対象を限定して索引付けを行った場合，ニュース映像が集中的に取り上げている時期の索引付けが疎になる可能性が高い．

3.3.2. 索引の補完

前節で示した問題を解決するために，文献[1]で紹介したトピックスレッド構造を利用して，索引付けの補完を行う．その概念図を図 4 に示す．図中の各ノードは，ストーリ分割後の各ニュース映像を表す．

トピックスレッド構造は強く関連するニュースを時系列に連鎖することにより構築されている．そのため，あるノードに対してスレッド構造上の前後に位置するノードは，意味的にも時間的にも最も類似するものであると考えられる．そこで，ある Wikipedia エントリが索引付けられているノードに対して，スレッド構造上の前後のノードに位置するニュースの CC とも類似度評価を行う．ここでしきい値を超えれば新たにそのノードにも当該エントリを索引付ける．この操作を類似度がしきい値以下になるノードが出現するか，既に同じエントリが索引付けられているノードが出現するまでトピックスレッド構造上で再帰的に適用する．既存の索引付けに対して以上の操作を繰り返すことにより，索引の補完を行う．

3.4. 各エントリに関連する映像群の抽出

索引付け結果を利用し，各 Wikipedia エントリに関連する映像群を抽出する．この際，索引付けられた各映像を放送日の早い順に並べる．索引付けられる映像がトピックスレッド構造上でクラスタを形成している場合，そのクラスタを保存し，提示する．

映像群の抽出・提示の際，文献[1]の手法を用いたトピックスレッド構造の再構築は行わなかった．これは，抽出される映像同士の放送日の間隔が離れることが多く，トピックスレッド構造のように分岐させて提示する必要性が低いためである．また，Wikipedia エントリとの対応付けにより話題が限定され，ストーリ群が分岐することが少なくなることも一因である．

4. 実験と考察

4.1. 使用するデータ及び実験条件

CC に関しては放送映像（NHK ニュース 7）に付随するものを使用した．2007 年 1 月 1 日から 2008 年 6 月 30 日までに放送された映像及び CC を使用し，CC はストーリごとに分割してあるものとする．また Wikipedia に関しては 2008 年 11 月 27 日付で記録されたデータベース・データをダウンロードして使用した．この時点での Wikipedia の全エントリ数は 1,053,561 件であった．また前章で説明した手法により抽出されたニュース関連エントリは 1,645 件であった．なお，この抽出の精度は十分良好であった．以降，4.2 節で対応付け精度の評価を行い，4.3 節で各エントリに対応付くストーリ数やそれらの日数の間隔を調査する．最後に 4.4 節で映像群の抽出結果の例を示す．

4.2. 実験 1　:対応付け精度

4.2.1. 実験条件・実験結果

各 Wikipedia エントリに対応付く CC を調査することにより，対応付け精度を評価した．評価対象は 3 個の Wikipedia エントリ（“新テロ対策特措法(*5)”，“大連立構想（日本 2007)”，“ねんきん特別便”）である．

対応付けの正誤は人手で判断した．ここで正しい対応付けとは，「CC 上で Wikipedia エントリに関する報道・説明を具体的に行っており，かつ Wikipedia エントリのテキスト中にもその説明が見受けられるもの」とした．適合率に関しては，対応付けられた CC 群の内容を全て調査し，人手で正誤判断を行った．また再現率に関しては，全ての CC の内容を調査することが困難なため，期間を 2～3 カ月に絞り，人手により正解データを作成し評価した．

日付情報により類似度評価の対象を限定するかどうか，さらに索引の補完を行うかどうかで比較実験を行った．手法 1 ：日付情報を利用せず補完もしない場合，手法 2：日付情報を利用するが補完は行わない場合，手法 3：日付情報を利用し補完も行う場合（提案手法）の 3 種類で実験した．実験結果を表 1 に示す．

表 1　実験結果：対応付け精度

	手法 1 日付× 補完×	手法 2 日付○ 補完×	手法 3 日付○ 補完○
適合率（%）	43.4	97.4	86.1
再現率（%）	95.1	45.4	79.3

＊5：正しくは“テロ対策海上阻止活動に対する補給支援活動の実施に関する特別措置法”

4.2.2. 考察

提案手法は適合率，再現率が共に高く，提案手法の有効性を確認した．適合率は手法 2 及び手法 3 が高かった．手法 1 は日付情報を考慮しないため，テキスト情報の類似度のみで対応付けられてしまい，誤対応が発生することが多かった．例えば“ねんきん特別便”というエントリでは年金に関する記述がなされており，年金について報道した日の CC と対応付けられることが多かった．しかし“ねんきん特別便”が報道される以前の，年金に関する問題を取り上げた日の CC とも対応付いており，これは適切ではない．他にもこのような誤対応が目立った．また再現率は手法 1，手法 3 が高かった．手法 2 の再現率は，Wikipedia エントリのテキスト中で日付情報が記述される頻度に大きく依存した．“新テロ対策特措法”に関する Wikipedia エントリでは日付情報がほとんど記述されておらず，再現率は 11%となった．しかし索引の補完を行うことにより，再現率は 63%に上昇した．

提案手法にも問題が見受けられた．まず索引の補完による適合率の低下である．トピックスレッド構造中で話題が少しずつ変化している場合があり，Wikipedia エントリとの関連度が小さくても，テキスト情報の類似度によって対応付いてしまうことが多かった．また再現率に関しても手法 1 より約 15%低い値となった．スレッド構造により全ての関連するニュースを網羅できていない場合があり，スレッド構造に依存しない対応付け手法も検討する必要があると考えた．

4.3. 実験 2 :ストーリ数と最大間隔

4.3.1. 実験条件・実験結果

各 Wikipedia エントリに対応付けられたストーリ数，及び対応付けられたストーリ群の中での最大間隔（日数）のそれぞれについて頻度を調査した．使用した Wikipedia エントリは，ニュース関連エントリとして抽出された 1,645 件である．実験結果を図 5 と図 6 に示す．なお双方ともストーリ数 0，最大間隔 0 日のデータは除去してある．なお，対応付けられたストーリ数が 0 である Wikipedia エントリは 1,305 件（79%），最大間隔が 0 日の Wikipedia エントリは 1,421 件（86%）であった．

4.3.2. 考察

各 Wikipedia エントリに対応付けられるストーリ数は 0～20 件であることが多かった．また，ストーリ数が 0 となるエントリが多かった最大の原因は，対応付けに使用したニュース映像の本数が少なく，かつ短い期間に放送されたものだけを使用したことであると考えられる．それに加え，NHK ニュース 7 は 1 回あたりの放送時間が短く，大きなニュースのみを取り上げることが多いことも原因の一つである．

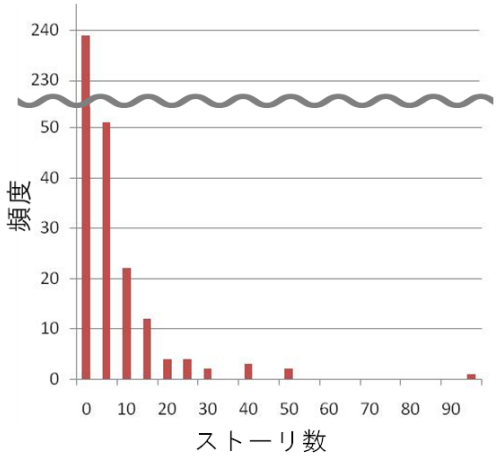

図 5　実験結果：ストーリ数の頻度

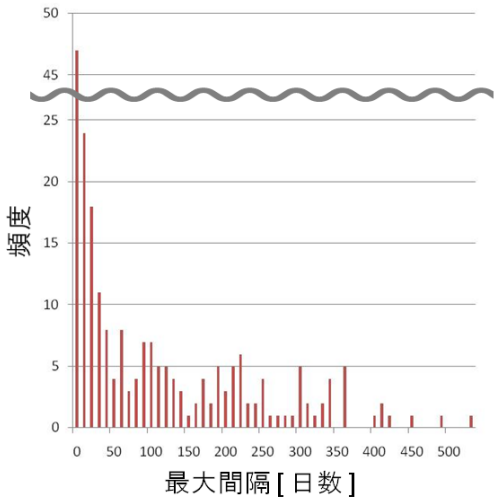

図 6　実験結果：最大間隔の頻度

最大間隔に関しては，300 日以上離れたストーリ同士が一つの Wikipedia エントリに対応付けられることもあり，放送日の離れた映像間の関連も発見できることを確認した．なお，図 6 のヒストグラム上で最大の階級値を示したエントリは誤対応であった．正しい対応付けとして最大の階級値を示したのは“NHK 番組改変問題”で，最大間隔は 498 日（2007 年 1 月 29 日～2008 年 6 月 10 日）であった．

4.4. 映像群の抽出結果

4.4.1. 実験条件・実験結果

提案手法を利用し，各 Wikipedia エントリに注目した映像群を抽出した．以下に“新テロ対策特措法”という Wikipedia エントリに関する抽出結果を示す．図 7 は 2007 年 9 月 8 日の番組の 3 番目のストーリを起点とするトピックスレッド構造の一部であり，各ノードは

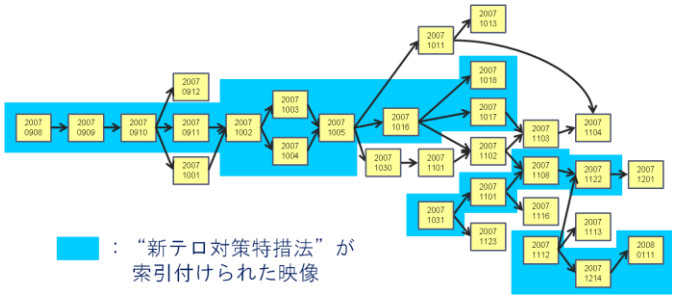

図 7　トピックスレッド構造の例

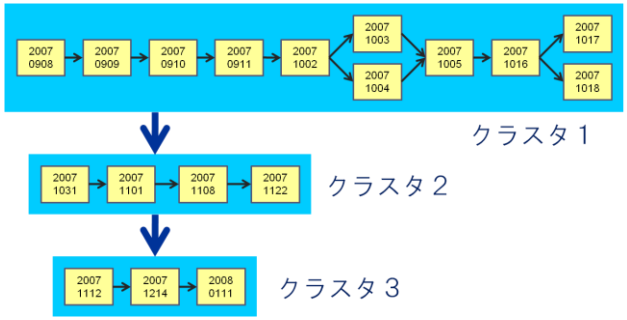

図 8　抽出された映像群の例

ストーリ分割後の各映像を表している．また“新テロ対策特措法”が索引付けられた映像を青色で示してある．図 8 はトピックスレッド構造から Wikipedia エントリに索引付けられた CC 群を抽出した結果である．

4.4.2. 考察

図 8 のように，抽出された映像群はいくつかのクラスタを形成する．ここで各クラスタにおいて，“新テロ対策特措法”と共に索引付けられた Wikipedia エントリを調査した．すると，クラスタ 1（2007 年 9，10 月）では“安倍改造内閣”など，クラスタ 2（2007 年 11 月）では“大連立構想（日本 2007)”など，クラスタ 3（2007 年 12 月前後）では“道路特定財源制度”などの Wikipedia エントリを確認した．これにより，各クラスタにおいて共起する Wikipedia エントリを分析することにより，各トピックスレッド構造中での話題の変遷を理解することができるのではないかと考える．

5. むすび

本稿では，Wikipedia を利用したニュース映像アーカイブの閲覧支援技術を提案した．実験により適合率 86%，再現率 79%で索引付けが行えることを確認した．また，放送日が離れた映像間の関連も発見可能であり，最大で 498 日の間隔がある映像同士を一つの Wikipedia エントリに対して対応付けることが可能であった．さらに，ある映像に共に索引付けられる Wikipedia エントリを分析することにより，話題の変遷の理解につながることを確認した．Wikipedia を利用することにより，クローズドキャプションのみによる処理では得られない情報を得られることを確認した．

今後の課題としてはまず，各 Wikipedia エントリに関連する映像群を直観的に提示する閲覧インタフェースを作成する．その際，索引付け精度の更なる向上を図る．また，Wikipedia エントリに対してだけでなく，テキスト中の各節や各文に対して詳細に映像群を対応付けることで，より直観的・効果的な閲覧技術を目指す．その他，話題の変遷を分析し提示する手法の検討や，ニュース映像以外の映像への適用を検討する．

謝　辞

実験データとして使用したニュース映像を提供して頂いた国立情報学研究所に感謝する．本研究の成果の一部は科研費による．本稿中の実験では SlothLib ライブラリ（http://www.dl.kuis.kyoto-u.ac.jp/slothlib）を使用しており，開発に携われた方々に感謝する．

文　献

[1] 井手一郎，木下智義，高橋友和，孟洋，片山紀生，佐藤真一，村瀬洋：“大量ニュース映像を対象とした時系列意味構造に基づく情報編纂手法の提案”，人工知能学会論文誌，Vol.23，No.5，pp.282－292 (Sep. 2008)

[2] National Institute of Standards and Technologies: “The year 2000 Topic Detection and Tracking (TDT2000) Task Definition and Evaluation Plan”, (2000), http://www.itl.nist.gov/iad/mig//tests/tdt/2000/

[3] P. Duygulu, J.-Y. Pan, and D. Forsyth：“Towards Auto-Documentary: Tracking the Evolution of News Stories”, Proc. 12th ACM Int. Conf. on Multimedia, pp.820－827 (Oct. 2004)

[4] X. Wu, C.-W. Ngo, and Q. Li：“Threading and Autodocumenting News Videos”，IEEE Signal Processing Mag., Vol.23, No.2, pp.59－68 (Mar. 2006)

[5] M. Rautiainen, T. Ojala, and T. Seppanen：“Cluster-Temporal Browsing of Large News Video Databases”，Proc. 2004, IEEE Int. Conf. on Multimedia and Expo, Vol.2, pp.751－754 (June 2004)

[6] C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, D. Koelma, G. Nguyen, O. de Rooij, and F. Seinstra：“MediaMill: Exploring News Video Archives based on Learned Semantics”，Proc. 13th ACM Int. Conf. on Multimedia, pp.225－226 (Nov. 2005)

[7] 中山浩太郎，原隆浩，西尾章二郎：“Wikipedia マイニングによるシソーラス辞書の構築手法”，情報処理学会論文誌，Vol.47，No.10，pp.2917－2928 (Oct. 2006)

[8] 川場真理子，中崎寛之，宇津呂武仁，福原知宏：“Wikipedia エントリとブログサイトの対応付けによる日本語ブログ空間のトピック分布推定”，情報処理学会研究報告，2008-NL-187 (pp.83－90) (Sep. 2008)