ニュース映像間の意味構造を利用した Wikipedia 情報の拡張

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† 岐阜聖徳学園大学 経済情報学部‡
国立情報学研究所§

1 はじめに

近年，Wikipedia はオンライン百科事典として重要な役割を担い始めている．しかし Wikipedia の記事には画像や映像，音声といったマルチメディア情報が付加されていることが少なく，それらが付加されたよりリッチな情報源を目指す必要がある．一方，放送やインターネットなどを通して提供される映像データは増加の一途をたどっている．中でも放送局により制作された放送映像は，アマチュアが制作した映像よりも高品質かつ信頼性が高いことが多いため，ビデオ・オン・デマンドサービスなどにより再利用されることが多くなっている．

このような背景を受け，例えば Miura らは，放送映像とそれらに付与された CC（クローズドキャプション；文字放送字幕）を利用し，マルチメディア百科事典の自動生成を行った[1]．一方，我々は本発表で，資料として価値の高いニュース情報に注目し，大量に蓄積されたニュース映像アーカイブ中の映像を利用した Wikipedia 情報の拡張方法を提案する．これにより各 Wikipedia エントリに対して，対応する放送映像を利用した詳細な説明の付加を実現する．映像と Wikipedia エントリとの対応付けは，映像に付与された CC と Wikipedia 中のテキストの類似度を評価することにより得る．また映像群を利用して Wikipedia 記事の閲覧を補助するインタフェースの試作についても報告する．

2 ニュース映像群と Wikipedia エントリの対応付け

2.1 手法の概要

処理の流れを図 1 示す．ニュース映像に関しては，映像に付与された CC を利用してトピックスレッド構造を構築しておく．この手法の詳細は文献[2]にゆずる．一方 Wikipedia に関しては，テキスト中から日付情報を抽出する．そして両方の出力結果から CC と Wikipedia エントリとの類似度評価を行い，対応付ける．この際，（1）各 Wikipedia エントリから抽出した日付情報により類似度評価の対象を限定した後，（2）トピックスレッド構造を利用し対応付けの拡張を行う．これらの工夫により対応付け精度の向上を図る．

Enhancing Wikipedia Information using the Semantic Structure between News Videos

† Tomoki OKUOKA, Daisuke DEGUCHI, Ichiro IDE, Hiroshi MURASE
Nagoya University, Graduate School of Information Science
‡ Tomokazu TAKAHASHI
Gifu Shotoku Gakuen University, Faculty of Economics and Information
§ Ichiro IDE
National Institute of Informatics

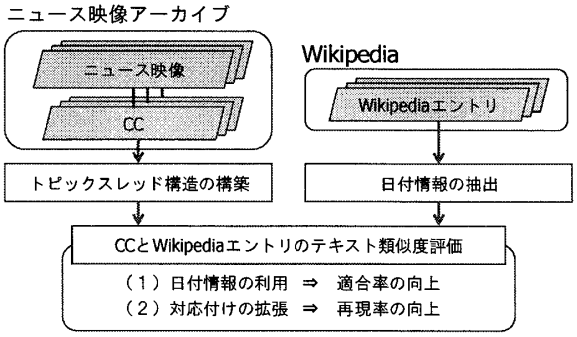

図 1　処理の流れ

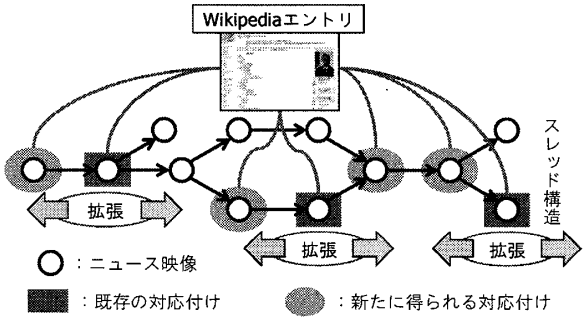

図 2　対応付けの拡張

2.2 テキストの類似度評価

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，しきい値を超えればそれらを対応付ける．

2.3 日付情報の抽出

全ての CC と Wikipedia エントリとの類似度評価を行うと，対応付け精度の低下を招く．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，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．テキスト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．前述の例では「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．

2.4 対応付けの拡張

日付情報により類似度評価の対象を限定した場合，あ

表 1　対応付け精度

	手法 1	手法 2	手法 3
テキスト類似度	✓	✓	✓
日付情報	–	✓	✓
対応付けの拡張	–	–	✓
適合率（%）	43.4	97.4	86.1
再現率（%）	95.1	45.4	79.3

るニュースイベントが注目されている時期において，対応付けの密度が粗くなることが多い．なぜなら，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のに対し，Wikipedia は，注目されている時期の日付情報を全て記述することは少ないためである．

以上の問題を解決するために，文献[2]の手法により得られるトピックスレッド構造を利用して，対応付けの拡張を行う．その概念図を図 2 に示す．図中の各ノードはニュース映像を表す．トピックスレッド構造上で，既に Wikipedia エントリと対応付いた映像の前後の映像に対し，再度類似度を評価する．この操作を再帰的に行い，対応付けを拡張する．

3 実験と考察

3.1 使用データ

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 テキストはトピック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 日付で記録された日本語版の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．

3.2 対応付け精度

各Wikipediaエントリに対応付くCCを調査することにより，対応付けの適合率及び再現率を評価した．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，期間を3カ月に絞り，人手により正解データを作成して評価した．日付情報により類似度評価の対象を限定するかどうか，さらに対応付けの拡張を行うかどうかで，以下の3種類の比較実験を行った．実験結果を表1に示す．

手法 1：日付情報を利用せず拡張も行わない
手法 2：日付情報を利用するが拡張は行わない
手法 3（提案手法）：日付情報を利用し拡張も行う

実験の結果，手法 3 は適合率・再現率が共に高く，提案手法の有効性を確認できた．

4 閲覧インタフェース

映像群を利用して各 Wikipedia 記事の閲覧を補助するインタフェースを試作した．その画面を図 3 に示す．画面左側には Wikipedia テキストが提示されている．なお，Wikipedia テキストは日付情報を利用して自動でブロック分割してある．また画面右側にはニュース映像が提示される．Wikipedia テキスト中の日付情報に対応したリンクをクリックすると，画面右側に対応するニュース映像が

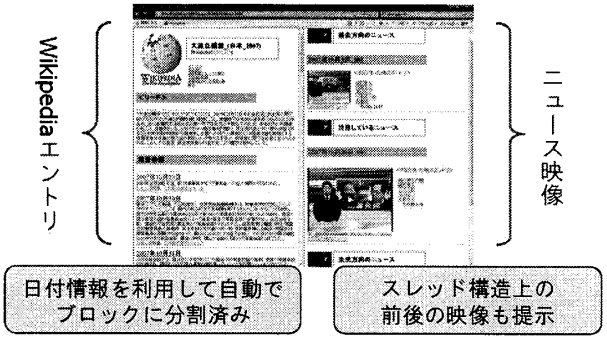

図 3　試作した閲覧インタフェース

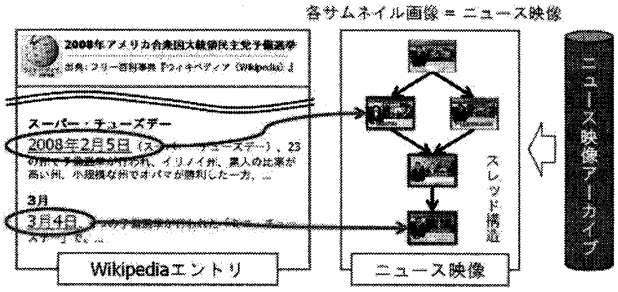

図 4　Wikipedia 記事のマルチメディア化

再生される．さらに各ニュース映像の上下には，トピックスレッド上の前後にあるニュース映像が提示されており，クリックすればそれらを閲覧できる．このようにスレッド構造上でニュース映像をたどることにより，Wikipedia テキスト中の出来事の経緯を，より詳細に理解することが可能となる（図 4）．

5 むすび

本発表では，ニュース映像アーカイブ中の映像を利用した Wikipedia 情報の拡張方法を提案した．この際，ニュース映像と各 Wikipedia 記事をテキスト情報の類似度評価により対応付けた．実験により適合率 86%，再現率 79%で対応付けが行えることを確認した．また，映像群を利用して各 Wikipedia 記事の閲覧を補助するインタフェースの試作についても報告した．

今後の課題としては，対応付け精度の更なる向上が挙げられる．また 1 つの Wikipedia エントリに対応付く映像数が大量である場合，映像の閲覧に時間がかかる．そこで，それらの映像を要約し，素早く閲覧させる技術にも検討する．

謝辞

実験に使用した映像を提供して頂いた国立情報学研究所に感謝する．本研究の一部は科研費による．

参考文献

[1] Miura et al., “Automatic Generation of a Multimedia Encyclopedia from TV Programs by Using Closed Captions and Detecting Principal Video Objects”, Proc. 8th IEEE ISM, pp. 873－880, Dec. 2006

[2] 井手ら，“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