ニュースアーカイブのための コンテンツ構成順序を用いた比較ニュース検索

北山　大輔† 　角谷　和俊††

† 兵庫県立大学大学院環境人間学研究科　〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
†† 兵庫県立大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
E-mail: †nd05r011@stshse.u-hyogo.ac.jp, ††sumiya@shse.u-hyogo.ac.jp

あらまし　現在，TV や新聞，インターネットなどを通して映像やテキストのニュースコンテンツが配信されている．一般にニュースは時間が経過すると価値が無くなると考えられる．しかし，現在閲覧しているニュースと関係するコンテンツであれば，過去のニュースであっても，同時に閲覧し比較することで，より理解を深めることが可能である．例えば，オリンピックなど何度も起こる類似のイベントにおける前回のメダル獲得時のニュースなどの場合である．そこで本研究では，ニュースアーカイブに対し，映像とテキストなど異メディアコンテンツの構成順序をもとに質問生成を行い，閲覧中のニュースをより理解するために比較ができるニュースコンテンツの検索方式を提案する．
キーワード　情報検索，情報統合，ニュースアーカイブ，ニュース映像

A Retrieval Method of Comparative News using Contents Structure Order for News Archives

Daisuke KITAYAMA† and Kazutoshi SUMIYA††

† Graduate School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
†† School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †nd05r011@stshse.u-hyogo.ac.jp, ††sumiya@shse.u-hyogo.ac.jp

Abstract　Video and text-news content have recently been broadcast on TV, newspapers, and the Internet. Although video content on out-ofdate news is of little value for viewing, it can be considered to have value by comparing it to related content. Repeated news should especially be compared, e.g., the Olympic games and international expositions. In that case, the more understanding might be deepened by comparing it. We propose a method of retrieving comparison content based on the order of news elements for news archives. It is composed of two parts. The first is analysis of news content that someone is browsing. The second is the automatic generation of queries for retrieving content on comparison news.
Key words　Infomation retrieval, Information integration, News archives, News video streams

1. はじめに

ニュースによる情報伝達は TV や新聞のみならずインターネットにおいても一般的となってきている．近年，映像ニュースはインターネット上でも FNN-NEWS.COM [3]，TBS News i [19]，日テレ NEWS24 [15]，ANN NEWS [1] といった各報道局により映像ニュースサイトが公開されてきている．しかし，映像ニュースサイトでは映像が公開されている期間は短く，長くても 1 週間程度である．また，テキストニュースもインターネット上の各社のウェブサイト (Sankei Web [17]，MSN Mainichi Interactive [11] など) で公開されている．これらテキストのニュースサイトでも，ニュースが公開されている期間は 1 ヶ月程度であり，期間が限定されている．これは，一般にニュースは速報性を重視しているためであると考えられる．

GoogleNews [4] において，News archive search というサービスが始まり，過去のニュースに対する関心が高まってきている．しかしながら，このサービスでは単純な検索機能とタイムラインにそって表示する機能という通常のニュース検索と同等の機能しか備えておらず，ニュースアーカイブを検索する方法として十分とはいえない．

一方，特集番組などの場合，過去のオリンピック競技の映像と現在の競技の映像を比較しながら映像を構成するということも行われている．そのため，ニュースとしての価値が失われた情報であっても，現在の事柄との関係性を示すことで閲覧する価値が生じるのではないかと考えられる．そこで本研究では，現在ユーザが閲覧しているニュースコンテンツに対し，比較することで現在のニュースをより深く理解できるようなニュースコンテンツの検索ができると便利であると考え，インターネット上で公開されているニュースコンテンツを収集しアーカイブするサイトにおいて，自動的に比較可能なニュースコンテンツを検索する手法を提案する．

以下，2 節において研究の概要と関連研究について述べ，3 節ではニュースコンテンツの特性に基づくキーワード抽出方法について説明し，4 節で比較ニュース検索のための質問生成方法を述べる．5 節でプロトタイプについて述べ，最後に 6 節でまとめと今後の課題について述べる．

2. 本研究の概要と関連研究

2.1 研究の概要

本稿では，ニュースコンテンツのメディアによる構成順序に基づき，ニュース中で対象となっている物事や，そのニュースカテゴリの種類を判定し，ユーザが現在閲覧しているニュースコンテンツと比較することが有効なニュースの検索を行う．

映像ニュースでは，撮影されている対象について述べるという特性から，その構成は，対象物単位であるものと考えられる．また，その内容の順序は，短時間で内容を的確に伝えるために時系列的に並びやすく,「発生 (Lead-in)，現状 (Body)，今後 (Standupper)」という順序で述べられることが多い [25]．一方，テキストニュースでは，概要に関する詳細，補足というように，ニュース全体を膨らませていく形で述べるという特性から，その構成は，ニュースの内容全体を単位としたものと考えられる．また，その順序も一般に,「概要，詳細，補足」というように，ニュースを理解するために必要な事象順 (逆ピラミッド型) に述べられることが多い [24]．このような特性に基づき，ニュースの対象・種類に関して抽出が行えると考えられる．

比較を行うことが有効なニュースは，対象もしくは種類の一方が同じであり他方が異なるような，ニュース中のいずれかを軸としているが，異なるニュースであると考えられる．そのため，抽出した要素を組み合わせることで比較ニュースを検索するための質問を生成することが可能であると考えられる．図 1 は本手法の概念図である．本手法の特徴は以下のとおりである．

ニュースメディアに基づくキーワード重要度算出　ニュースコンテンツは，そのメディアに応じてもっとも適切に情報伝達されるように構成がされていると考えられる．そこでニュースメディアに応じた，キーワード重要度算出を行う．ニュースで述べられている対象を表すキーワードに関しては，その構成単位に基づき重要度算出を行い，そのニュースの種類を表すキーワードに関しては，その出現順序に基づき重要度算出を行う．

コンテンツ構成グラフに基づく質問生成　コンテンツを構成する要素として “対象”，“種類” を考え，コンテンツ自身の構成を

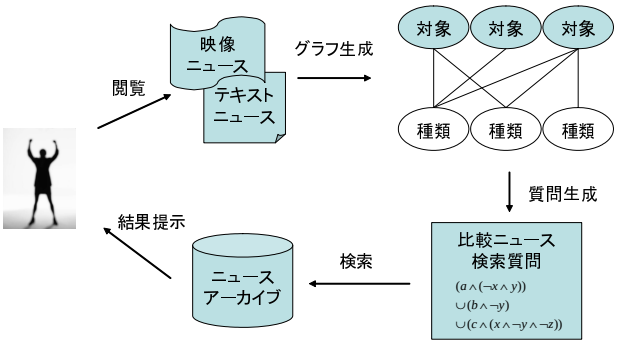

図 1　比較ニュース検索の概念図

表現するために，対象と種類の対応関係を定義する．対応関係もメディアによって異なり，キーワードの出現範囲に基づき対応付けを行う．質問生成は，コンテンツ構成グラフとキーワード重要度を用いて行い，用いるキーワードは，コンテンツ構成グラフの部分グラフに出現するキーワードである．

比較ニュース検索　比較ニュース検索として “対比質問” による検索と “類比質問” による検索を定義する．対比質問は，対象を軸として比較が行えるニュース記事を検索する質問であり，類比質問は，ニュースの種類を軸に比較が行える記事を検索する質問である．キーワード重要度とコンテンツ構成グラフにより，キーワードを AND,OR,NOT で接続することで比較検索を可能とする．

このことにより，ユーザはニュースコンテンツを閲覧し比較の種類を選択するだけで，自動的にニュースをより理解するためのコンテンツを得ることが可能となる．

2.2 予 備 実 験

同一のニュースを扱っていたとしても，メディアによってその表現の手段が異なると考えられる．テキストのニュースであれば，時系列に関係なく読み返すことができるため，論理的に説明するように構成されていると考えられる．また，映像のニュースは映像の撮影対象に依存するため，ある対象はある区間に集中して出現する内容構成になると考えられる．各メディアによるニュースの構成のされ方の特性を明確にするために予備実験を行った．実験に用いた映像ニュースは FNN-NEWS.COM，TBS News i，ANN NEWS，日テレ NEWS24，テキストニュースは Sankei Web，MSN 毎日インタラクティブ，asahi.com，Yomiuri Online を用いた．FNN-NEWS.COM は Sankei Web に対応するなど，それぞれの記事が対応している考えられ，これらのニュースサイトより対応する 14 件の記事を選択し，実験を行った．対応していると考えられるニュース同士を比較しているのは，映像とテキストのメディアによる違い以外，例えば報道スタンスの違いなどによる構成の差を可能な限り減らすことができると考えたためである．手順は以下のとおりである．

1. 映像ニュースの音声テキストより，文(注1)を時系列出現順に並べ，内容構成についての特徴を考察する．

(注1)：本稿では，主語・述語が含まれているかにかかわり無く，読点で区切られた範囲を文と呼ぶ

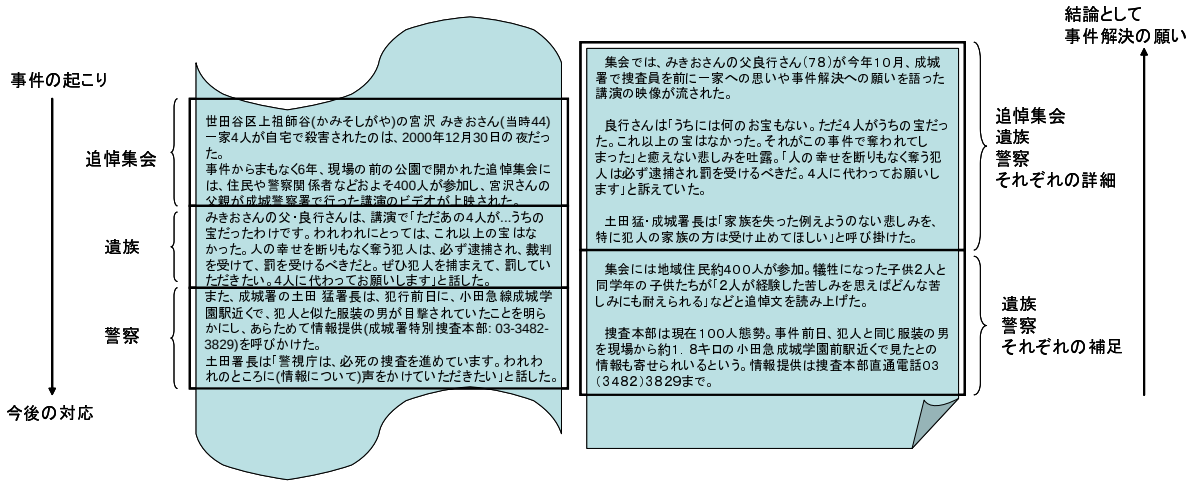

図 2　予備実験：世田谷一家殺害事件追悼集会の例

2. テキストニュースより，文を文書内出現順に並べ，内容構成についての特徴を考察する．
3. 映像ニュースとテキストニュースの各文の対応関係をつけ，構成順序に関しての考察を行う．

例として取り上げたニュースを図 2 に示す．結果は以下のとおりである．

映像の内容構成がある対象に関する部分の組み合わせによって成り立っていると判断できたニュースは 11 件であった．図 2 では，ニュースのあらまし，被害者家族の話，警察関係者の話といった部分から成り立っていると判断できる．判断できなかったニュースは，交通違反のニュースや，架空登録のニュースなど，明確に対象がわからないニュースであり，映像も現場風景のような抽象的な場面で構成されていたものである．

テキストの構成が概要で述べられた内容を中心に，詳細，補足という単位の部分から成り立っていると判断できたニュースは 12 件であった．図 2 では，始めの概要に対し，被害者と警察のコメントという詳細，それに対する地域住民の対応と警察の対応という補足から成り立っていると判断できる．映像における構成では，1 箇所であった警察に関する内容が 2 個所になっており，映像とテキストの構成基準の差が現れていると考えられる．判断できなかったニュースは，交通違反のニュースや，捜査方針を切り替えたというニュースなど，詳細に述べるために書かれたのか，補足的に書かれたのかが不明確であるニュースであった．

映像の構成順序とテキストの構成順序が異なるニュースは 14 件であった．その構成の順序の違いの共通点として，映像では時系列順，テキストではまとめを導く過程の逆順であると考えた．実際に，映像ニュースで時系列順に述べられていると判断できたニュースは 10 件であり，テキストニュースで，概要の直後にまとめに当たる内容が述べられていると判断できたニュースは 11 件であった．図 2 の映像側では，「事件の起こり，今回の追悼式，警察の今後の対応」という時系列順であると判断でき，テキスト側では「事件解決の想い，追悼式の様子，警察の情報」という，「事件解決の想い」を導くための情報が，後に出現しているような順序であると判断できる．

これらのことより，以下のことを確認した．

- 映像ニュースの構成単位は映像内の対象物であり，ある対象はそのシーンを中心に述べられる．
- テキストニュースの構成単位は概要で述べられた内容であり，展開するに従い，ニュース中の対象物が何度も述べられる．
- 映像ニュースは時系列順に構成されやすく，終わりに近づくほど今後の展開を述べる
- テキストニュースは結論順に構成されやすく，はじめにこのニュースの展開を述べる．

本稿では，これらの特性を用いてニュースの構成を抽出し，その構成が部分的に異なる比較ニュースの検索質問生成が行えるものと考えた．

2.3 関 連 研 究

現在提供されているニュースサイトとして McKeown ら [10] の Newsblaster [13] や Radev ら [16] の NewsInEssence [14] や GoogleNews があげられる．これらのニュースサイトは主として，そのトピックを簡潔に理解するための続報記事の集約・要約を目的としており，本稿で提案するトピックを問わない，比較可能なコンテンツの検索とは目的が異なる．

ニューストピックを俯瞰的に見る研究として渡邊ら [21] や井手ら [6]，吉岡ら [22] の研究があげられる．これらの研究は，あるニュースを多面的に見るという点で本研究と類似している．しかしながら，従来の研究は同一トピック内のニュースにとどまっており，異トピックのニュースを用いてあるニュースを多面的に見る本手法とは異なる．

ニュースの構成要素の抽出に関する研究として井手ら [7] や戸田ら [20] の研究があげられる．井手らの手法はニュース映像のオープンキャプションやクローズドキャプションよりニュースの構成要素として 4W(Who,Where,When,What) に相当するキーワードの抽出を行うものである．戸田らの手法はニュースの構成要素として固有表現に着目し，トピックごとに固有表現の種類に対する重みを変化させることによりクラスタリングの精

度を向上させる手法である．本手法は，このような特定のキーワードに依存してニュースの構成を抽出するのではなく，複雑な文法解析や辞書構築などを必要とはしない．

複数のコンテンツ間の関係を求める研究として，張ら [23] や灘本ら [12]，北山ら [27] の研究があげられる．張らは，長期間続いているトピックに対し，意味的に影響を与えたトピックを求める手法を提案している．本手法は，時系列データのパターン解析により関係を抽出するのではなく，単一ニュースの構造のみを用いて関係を求める点で異なる．灘本らや北山らは，コンテンツの文書ベクトルの関係からコンテンツ間の関係を導く手法を提案している．本手法は，ベクトル空間を用いずに，キーワードベースで関係を導くという点で手法が異なる．

あるコンテンツから検索質問を生成する研究として Henzinger ら [5] や馬ら [26] の研究があげられる．Henzinger らはニュース映像から自動的に質問を生成し，その内容に類似した Web ページを検索する手法を提案している．馬らはテレビ番組から自動的に，内容を幅広くカバーするための質問や掘り下げるための質問など，数種類の質問を生成しテレビ番組の内容を補完できる Web ページを検索する手法を提案している．本研究は，内容の補完や補足を行うのではなく，あるニュースを別のニュースと比較することで内容の理解を深めるということを目的としている点で異なる．

3. コンテンツ構成順序を用いたキーワード抽出

3.1 コンテンツ構成順序と比較ニュース

コンテンツ構成順序とは，一つのニュースコンテンツの構成のされ方とその順序であり，メディアによってそれぞれ異なる特徴を持つ．本手法では，ニュースコンテンツの内容構成を用いて主体となっている対象を抽出し，内容順序を用いてニュースカテゴリの種類を抽出する．図 3 はコンテンツ構成順序の模式図である．図中の個々のシーンが構成要素であり，シーンの並びが構成順序である．また，本方式では，概要部分は扱わない．

テキストニュースであれば，概要に対する詳細情報を述べ，その後に補足情報が続くといった構成がされている．それに対し，映像ニュースでは，撮影されている対象に対し，詳細も補足もまとめて述べるという構成がされる．つまり，テキストニュースでは，ニュース全体を単位として，概要・詳細・補足と展開し，映像ニュースでは，撮影対象を単位として，その対象の詳細・補足と展開する．また，テキストニュースでは，ニュースの理解に必要な順序で内容が述べられるのに対し，映像ニュースでは，主として時系列に内容が展開するように内容が述べられる傾向にある．

比較ニュースとは，あるニュースに対し対象に着目して比較を行うことができるニュースや，種類に着目して比較を行うことができるニュースのことであり，前者を対比ニュース，後者を類比ニュースと呼ぶ．なお，続報のように連続的な事柄の前後を提示して比較するといったものや，異なる報道局の同一ニュースを提示して比較するということも考えられるが本研究では扱わない．

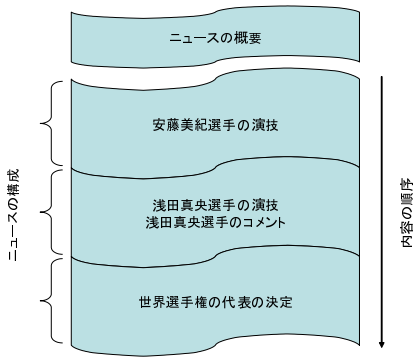

図 3　ニュース構成順序

比較可能なニュースコンテンツを検索する場合には，そのニュースで述べられている対象とニュースの種類を抽出する必要がある．ニュースの対象は名詞で表現されていることが多く，ニュースの種類は特定の動詞に現れていると考えられる．例えば,「小泉首相が退任した」というようなニュースであれば，“小泉首相” という対象と，“退任する” というカテゴリの種類を現す表現が使われるようなニュースであると考えられる．また,「田中知事が退任した」というニュースであれば，“田中知事” が対象であり，“退任する” という表現が使われるニュースであるといえる．このようにニュースカテゴリの種類が同じであれば同じ動詞が使われていると考えられる．また，一つのニュースは，対象として名詞，種類として動詞によって表すことが可能であると考えられる．

3.2 映像ニュースからのキーワード重要度算出

映像ニュースからの質問生成のためのキーワード重要度の算出について説明する．まず対象を表すキーワード重要度の算出について述べ，次に種類を表すキーワード重要度の算出について述べる．映像ニュースにおいて，ニュースの対象を表すキーワードは，対象が撮影されているシーンに名詞として集中して出現することが考えられる．例えば,「小泉首相が靖国参拝をした」というニュースであれば，いずれかのシーンにおいて，映像中に小泉首相が出現し，その前後で “小泉首相” という名詞が頻出することが考えられる．このような特徴から，映像ニュースの音声テキスト中の単語密度により対象を表すキーワードとして名詞の重要度算出を行う．図 4 左側に映像ニュースの対象重要度算出を示した．名詞 a の重要度は以下の式により算出を行う．

$$obj\_val = \frac{n}{dist(a_1, a_n)} \tag{1}$$

式中の a_n は n 番目に出現する名詞 a であり，$dist$ 関数により文距離を算出する．文距離は，何文離れているかを表す数であり，同一文中に出現する場合を 1 とする．この式により，単語の出現区間に何回出現するかという密度を算出し，この値が大きいほどニュース中での対象として述べられている可能性が高いものとする．

映像ニュースにおいて種類を表すキーワードは，今後の展開を述べる映像の終端付近に動詞として出現しやすいと考えられる．映像ニュースの内容順序の特徴として，まず，時系列的に

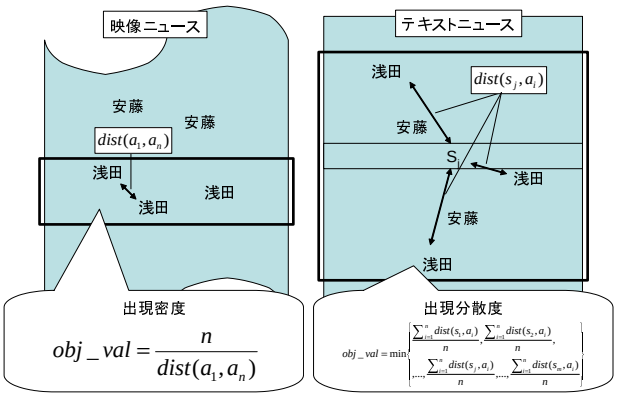

図 4　対象重要度算出

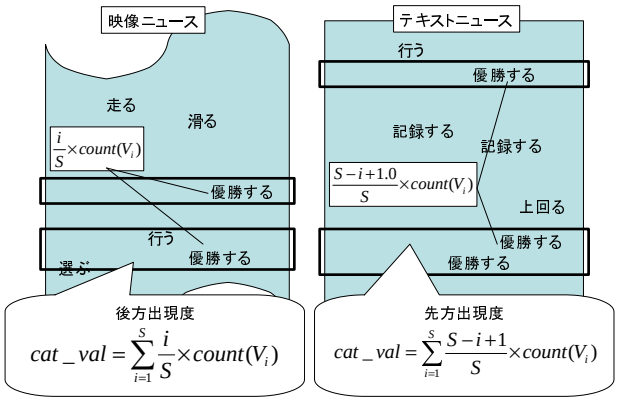

図 5　種類重要度算出

“何がおきた” という過去のことを述べ，“どのようになっている” という現在のことを述べ，最後に “今後どのようになる” ということを述べる．つまり，終端がまとめにあたると考えられ，まとめ部分での動作を示す動詞がニュースの種類を表すと考えられる．このような特徴から，映像ニュースの音声テキスト中の出現箇所により種類を表すキーワードとして動詞の重要度算出を行う．図 5 左側に映像ニュースの対象重要度算出を示した．ある動詞の重要度は以下の式により算出する．

$$cat\_val = \sum_{i=1}^{S}(\frac{i}{S} \times count(V_i)) \quad (2)$$

式中の i は S 文中の i 番目の文であることを表し，$count$ 関数により，i 番目の文に出現する動詞集合 V 中における重要度算出対象の動詞の個数を算出する．この式により，映像ニュースの音声テキストにおける文の位置による重要度を算出し，この値が大きいほどニュースの種類を表すキーワードの可能性が高いものとする．

3.3　テキストニュースからのキーワード抽出

テキストニュースからの質問生成のためのキーワード重要度の算出について説明する．まず対象を表すキーワード重要度の算出について述べ，次に種類を表すキーワード重要度の算出について述べる．テキストニュースにおいて，ニュースの対象を表すキーワードである名詞は，一箇所に集中して現れることは無いと考えられる．例えば,「小泉首相が靖国参拝をした」というニュースであれば，ある部分では,「小泉首相の靖国参拝」に関しての詳細を述べ，ある部分ではその補足を述べる．というように，対象を表す名詞はニュース記事内でさまざまな箇所に出現すると考えられる．このような特徴から，テキストニュース記事中の単語分散度により対象を表すキーワードとして名詞の重要度算出を行う．図 4 右側にテキストニュースの対象重要度算出を示した．名詞 a の重要度は以下の式により算出を行う．

$$obj\_val = min\Big(\frac{\sum_{i=1}^{n} dist(s_1, a_i)}{n}, \frac{\sum_{i=1}^{n} dist(s_2, a_i)}{n}, \ldots, \frac{\sum_{i=1}^{n} dist(s_j, a_i)}{n}, \ldots, \frac{\sum_{i=1}^{n} dist(s_m, a_i)}{n}\Big) \quad (3)$$

式中の a_i は i 番目に出現する名詞 a であり，s_j はテキストニュース内の j 番目の文である．$dist$ 関数により文距離を算出する．文距離は，何文離れているかを表す数であり，同一文中に出現する場合を 1 とする．min 関数により，要素中の最小値を抽出している．min 関数を用いるのは，単語分散の期待値が不明であるため，分散度が最も低くなる位置を最適な期待値として値を求めるためである．この式により，テキストニュース中の単語の分散度合いを算出し，この値が大きいほどニュース中での対象として述べられている可能性が高いものとする．

テキストニュースにおいて種類を表すキーワードである動詞は，ニュースのまとめ部分を詳細に述べているテキストの開始部分付近に出現しやすいと考えられる．テキストニュースの内容順序の特徴として，ニュースの理解に重要なことから先に書かれていると考えられる．つまり，始端がまとめにあたると考えられ，まとめ部分での動作を示す動詞がニュースの種類を表すと考えられる．このような特徴から，テキストニュースの記事中の出現箇所により種類を表すキーワードとして動詞の重要度算出を行う．図 5 右側にテキストニュースの種類重要度算出を示した．ある動詞の重要度は以下の式により算出する．

$$cat\_val = \sum_{i=1}^{S}(\frac{S-i+1}{S} \times count(V_i)) \quad (4)$$

式中の i は S 文中の i 番目の文であることを表し，$count$ 関数により，i 番目の文に出現する動詞集合 V 中における重要度算出対象の動詞の個数を算出する．この式により，テキストニュースの記事中における文の位置による重要度を算出し，この値が大きいほどニュースの種類を表すキーワードの可能性が高いものとする．

4.　比較ニュース検索のための質問生成

4.1　コンテンツ構成グラフの生成

質問の生成は，コンテンツ構成を表現するグラフを用いて行うする．コンテンツ構成グラフとは，対象と種類の重要度を持つキーワードからなる二項グラフであり，そのリンクは対象と種類の対応関係を表す．コンテンツ構成グラフにより，ニュースが出現するキーワードによってどのように構成されているかを表現することができる．

対応の決定はニュースのメディアによって異なり，映像ニュースでは，ある対象キーワードに対応する種類キーワードは，対

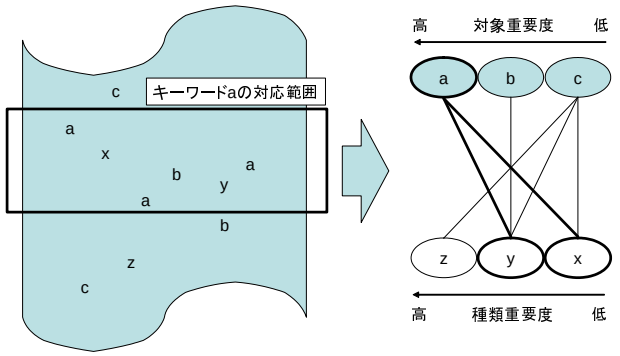

図 6　コンテンツ構成グラフ

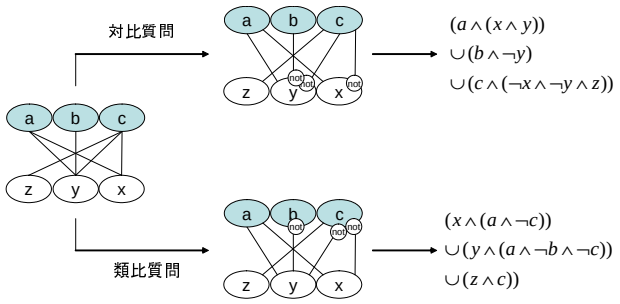

図 7　比較質問生成

象キーワードの出現密度の高い範囲において出現すると考えられ，その範囲に出現する種類キーワードとリンクを形成する．テキストニュースにおいては，同一パラグラフに出現する対象キーワードと種類キーワードが対応すると考えられ，同一パラグラフにおける対象キーワードと種類キーワードでリンクを形成する．図 6 に映像ニュースの場合のコンテンツ構成グラフの例を示した．図では，対象キーワード a に対して，その範囲内に含まれる種類キーワード x，y が対応づけられる様子を示している．また，コンテンツ構成グラフでは，左のキーワードから重要度順に表示するものとする．

4.2　対比質問生成

現在見ているニュースの対比ニュースを検索するために，ニュース内のキーワードを用いて自動的に質問を生成する．対比ニュースの検索は，現在見ているニュースに対し，ニュースで述べられている対象は同じであるが，その種類が異なるニュースを抽出することによって行う．例えば，“小泉首相の国会答弁” であれば，対比ニュースとして “小泉首相の応援演説” というように，“小泉首相” という対象に関して，“応援演説” という種類の異なるニュースを得ることで，普段から一貫した主張をする人物なのかという確認を行うことができる．

対比質問の生成は以下の手順により行う．

1. ある対象とリンクしている種類は AND 条件で接続する
   - 対象重要度の高い対象と接続している種類重要度の高い種類は接続の際に NOT 条件とする
   - 対象重要度の低い対象と接続している種類重要度の低い種類は接続の際に NOT 条件とする
2. 同じ対象と，接続している種類からなる質問の接続を行う
   - NOT 条件ではない種類からなる質問を OR 条件で接続する
   - NOT 条件の種類からなる質問を OR 条件で接続する
3. 2. で生成された対象重要度が閾値以上の質問による検索結果を OR 条件で結合する

対比質問により，いくつかの対象に対して，それぞれの現在見ているニュースとは異なる種類のニュースが検索結果として得ることができる．対象重要度の閾値を調節することで，対比関係の度合いを調節することができる．閾値を上げることで，対比関係として絞り込むことが可能となる．対比質問の生成の様子を図示したものが図 7 の上部である．

4.3　類比質問生成

現在見ているニュースの類比ニュースを検索するために，ニュース内のキーワードを用いて自動的に質問を生成する．現在見ているニュースに対し，ニュースで述べられている対象は異なるが，その種類が同じニュースの検索を行う．例えば，“ライブドアのニッポン放送買収” であれば，類比ニュースとして “楽天の TBS 買収” というように，“買収” という種類に関して，“楽天”，“TBS” という種類の異なるニュースを得ることで，現在見ているニュースにおける “買収” というものがどのような位置づけであるのかを比較することができる．

類比質問の生成は以下の手順により行う．

1. ある種類とリンクしている対象は AND 条件で接続する
   - 種類重要度の高い種類と接続している対象重要度の高い対象は接続の際に NOT 条件とする
   - 種類重要度の低い種類と接続している対象重要度の低い対象は接続の際に NOT 条件とする
2. 同じ種類と接続している対象からなる質問の接続を行う
   - NOT 条件ではない対象からなる質問を OR 条件で接続する
   - NOT 条件の対象からなる質問を AND 条件で接続する
3. 2. で生成された質問のうち種類重要度が同じ質問を OR 条件で接続する
4. 生成された種類重要度が閾値以上の質問を AND 条件で接続する
   - 重要度が高い種類を含む質問を接続する
   - 重要度が低い種類を含む質問は接続しない

類比質問により，現在見ているニュースの種類に対して，ニュースの対象が異なるニュースが検索結果として得ることができる．種類重要度の閾値を調節することで，類比関係の度合いを調節することができる．閾値を下げることで，類比関係として絞り込むことが可能となる．類比質問の生成の様子を図示したものが図 7 の下部である．

5.　評　　価

5.1　プロトタイプシステム

プロトタイプシステムは大きく分けて，ニュース構成順序解

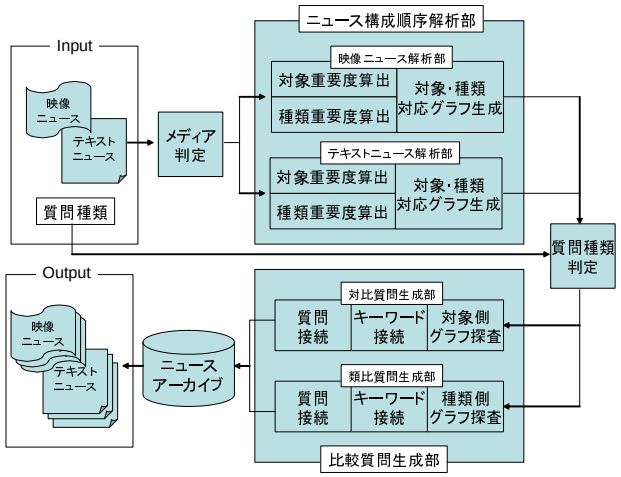

図 8　システム構成図

図 9　プロトタイプ画面

析部と質問生成部の 2 つの部分からなる (図 8)．ニュース構成順序解析部では，ニュースコンテンツが映像なのか，テキストなのかというメディア特性に基づき，対象・種類それぞれのキーワードの重要度を算出している．質問生成部では，キーワード重要度と対象・種類の対応関係から指定された種類の検索質問を生成し，ニュースアーカイブに対して検索を行っている．

図 9 はプロトタイプシステムの画面イメージであり，手前に表示したものが類比検索の例，他方は同じニュースで対比検索をした時の例である．ユーザは画面左上の現在閲覧しているニュースコンテンツに対し，知りたい比較ニュースの種類を左下のボタンで選択するだけで，システムが自動的に検索を行い，検索結果を画面右下のリストボックスに表示する．ユーザは，検索結果を選択することで比較ニュースを得ることができる．また，生成されたクエリを左下のボタン上のテキストボックスに表示するため，どのような条件で検索された結果なのかを知ることが可能である．現在見ているニュースサイトを切り替えたい場合は，左側のリストボックスから選択することで自由にニュースサイト間を行き来することが可能である．

プロトタイプで用いたニュースアーカイブは，FNN-NEWS.COM，TBS NEWS i の 2 種類の映像ニュースサイト，

表 2　キーワード重要度算出の結果

	適合率	
	対象	種類
テキスト	0.53	0.56
映像	0.68	0.27

Sankei Web，MSN Mainichi Interactive の 2 種類のテキストニュースサイトを 1 年 6ヶ月アーカイブしたものであり，ネットワーク上に配置している．ニュースアーカイブ内の検索は Interstage Shunsaku Data Manager [8] を用い，構成順序解析部，質問生成部はともに Visual Studio 2005 の C# により作成した．ニュースコンテンツからの単語抽出には SlothLib [18] を用いた茶筌 [2], [9] による形態素解析を用い抽出した．

5.2　キーワード重要度算出の精度

本手法のニュース構成順序に基づくキーワード重要度の判定を評価するために実験を行う．あるニュースに対して被験者が対象・種類に相当するキーワードの判定を行い，そのキーワードを解とした適合率で評価を行う．実験に用いたニュースを表 1 に示す．実験に参加した被験者は 5 人である．ニュースは，同日に対応するサイトで同様の内容で報道されたニュースという基準で取得して用いた．

被験者は，ニュースのタイトルおよび概要の部分を知らされずに，内容のみを見て解答した．同日かつ同トピックの他メディアのニュースとセットにして評価をしているのは，メディアによる差異を明らかにするためである．表中の V は映像ニュースであることを示し，T はテキストニュースであることを示している．

実験の手順は以下の通りである．

1. ニュースコンテンツより，提案手法を用いて対象・種類の重要度付けを行う
2. 被験者がニュースコンテンツを閲覧し，何を対象に評価を述べているかを判定し，キーワード 3 語を選択する
3. 同様に，ニュースの種類を特徴付けている一文を選択し，その文中の動詞を抽出する
4. システムの算出したキーワードに関して重要度上位 40%を抽出し，被験者の選択したキーワードと一致する割合により適合率を算出する

評価実験の結果を表 2 にまとめた．また，個々のニュースにおける適合率は表 1 に併記した．以下に考察を行う．

- 映像における種類適合率が低い結果となっている．これは，映像ニュースの構成のされ方が映像ニュース自身の長さに左右され，短いニュースでは本研究の仮説と異なる構成順序となっているためであると考えられる．ニュースの構成のされ方により，被験者の選択した正解が前半に出現する文のキーワードであることも多く，ニュースの長さによって抽出方法を変更するといった対処が考えられる．
- 双方のメディアの対象重要度の精度がおおむね同程度という結果となった．これは，メディアの特性を考慮した

表 1　キーワード重要度算出の実験データ

	タイトル	適合率 対象	種類
V	がんに効くとされる「アガリクス」含む健康食品から発がん促す作用検出　販売中止決定 2006/2/14	0.50	0.60
T	アガリクス製品を販売中止　「発がん促進する疑い」 2006/2/14	0.50	0.57
V	元スパイ殺害　英大使館から放射性物質痕跡 2006/12/7	0.60	0.50
T	リトビネンコ元中佐の不審死，殺人と断定…英警察 2006/12/7	0.60	1.00
V	大阪市姉妹殺害強盗放火事件　山地悠紀夫被告に死刑判決 2006/12/13	0.67	0.09
T	大阪の姉妹刺殺，放火の男に死刑「冷酷で非道，反省なし」 2006/12/13	1.00	0.38
V	交通違反の逃走車に警官が発砲 2006/12/10	0.67	0.75
T	パトカーに車ぶつけ逃走図る，警官発砲し逮捕…北九州 2006/12/10	0.80	0.43
V	ロシア元スパイ不審死事件　ロンドン警視庁，殺人事件と断定し捜査との声明 2006/12/7	0.40	0.33
T	在露英大使館からも放射性物質　英警察は殺人で捜査中 2006/12/7	0.50	0.50
V	謝罪の言葉なし…姉妹殺害で山地被告に死刑判決 2006/12/13	0.83	0.38
T	大阪・浪速の姉妹殺害，被告に死刑判決　大阪地裁 2006/12/13	0.17	0.00
V	世田谷一家 4 人殺害事件からまもなく 6 年　警視庁，追悼集会を開き情報提供を呼びかけ 2006/12/10	0.60	0.00
T	「4 人は宝」と遺族訴え　世田谷一家殺害で追悼集会 2006/12/10	0.33	0.67
V	ライブドア事件　東京地検特捜部，堀江貴文前社長らを証券取引法違反の罪で起訴 2006/2/14	0.80	0.20
T	きょう堀江容疑者ら起訴　粉飾決算容疑で再逮捕へ 2006/2/14	0.50	0.45
V	自民党　造反組復党で公認調整は先送りの方針 2006/12/5	1.00	0.14
T	刺客 6 人，中川幹事長から経緯説明「少しほっと」 2006/12/5	0.67	1.00
V	大阪市西成区の 40 平方メートルの建物に 3300 人もの住民登録大半が架空登録か 2006/12/7	0.75	0.17
T	西成 5 階建てビルに 3300 人住民登録給付金目的か 2006/12/7	0.67	0.33
V	「アース製薬」未公開株が流出　株販売の名古屋市内の無登録業者などを家宅捜索 2006/2/14	0.75	0.38
T	アース製薬未公開株を違法販売　無登録業者を家宅捜索 2006/2/14	0.80	1.00
V	貧困対策の経済学者ユヌス氏にノーベル平和賞を授与 2006/12/10	0.20	0.53
T	ユヌス氏「貧困は平和への脅威」ノーベル平和賞授賞式 2006/12/10	0.50	0.00
V	都知事主催イベント製作の四男旅費を都負担と共産党 2006/12/6	0.50	0.00
T	契約書類に「知事四男」の文字なし　ダボス公費渡航 2006/12/6	0.33	0.54
V	大阪の姉妹殺害事件で被告に死刑判決 2006/12/13	0.50	0.10
T	大阪姉妹殺害　山地被告に死刑判決　大阪地裁 2006/12/13	0.33	0.50

算出手法が，同じ尺度で重要度計算を行えているためであるといえる．異メディアを対等に検索するためのキーワード重要度判定手法として用いることができる可能性があると考えられる．

- 本手法では,「事件」や「事故」などの一般的なキーワードも重要度が高くなる場合がある．例えば,「世田谷一家 4 人殺害事件の追悼集会」のニュースなどで，それらの重要度が高くなっていた．このような場合では,「ある事件から一定期間が経過した」ような比較ニュースを検索するためのキーワードとしては適切な重要度付けを行えているといえる．被験者の判定でも，このニュースにおいて対象として「事件」というキーワードを選択したケースがあった．

これらのことより，ニュースのメディアの構成順序に基づくキーワードの重要度付け手法により，各メディアともに適切な尺度でキーワード重要度を算出できる可能性があるといえる．しかしながら，精度として十分であるとはいえず，アルゴリズムの改良や，他の手法との比較実験などが今後の課題といえる．

5.3　比較ニュース検索の精度

提案手法のコンテンツ構成順序を用いた質問生成の検索結果に関する評価実験を行う．実験データごとに，データセットとして 180 件程度の記事を用いた．(注2) この実験用データセットには，テキストニュースも映像ニュースも含まれている．

(注2)：180 件としたのは，18ヶ月のデータに対し，1 月に 8 記事ずつサンプリングを行ったものに，実験のために 40 記事程度正解の候補を入れたためである．

表 4　比較ニュース検索の結果

ニュース番号		対比質問 適合率	再現率	F 値	類比質問 適合率	再現率	F 値
1	テキスト	0.50	0.25	0.33	0.80	0.33	0.47
	映像	1.00	0.25	0.40	0.64	0.39	0.49
	All	0.67	0.25	0.37	0.68	0.37	0.48
2	テキスト	0.00	0.00	0.00	0.43	0.33	0.38
	映像	0.00	0.00	0.00	0.44	0.71	0.54
	All	0.00	0.00	0.00	0.43	0.57	0.49
3	テキスト	0.00	0.00	0.00	0.50	0.25	0.33
	映像	0.33	0.40	0.36	0.00	0.00	0.00
	All	0.22	0.18	0.20	0.27	0.14	0.19
4	テキスト	0.29	0.40	0.33	0.05	0.33	0.08
	映像	0.17	0.50	0.25	0.33	0.27	0.30
	All	0.20	0.43	0.27	0.13	0.29	0.18

表 3 に実験に用いたタイトル，メディア，生成された質問を記載した．実験データは，人手で見て適切にキーワード重要度付けが行えていると判断できたものを映像とテキストで同数用いた．データセット中より，各検索質問の種類ごとに，被験者が正解を抽出した．実験は，質問を生成したニュース記事と検索質問の種類のみを提示した状態で，データセット中の記事を一つずつ閲覧して行った．被験者の人数は 3 人 1 組で行い，2 人以上の被験者が正解とみなした記事を正解の記事とした．この実験の評価は，データセット中の正解に対する適合率，再現率，F 値で行った．実験の手順を以下に示す．

1. システムが対比質問，類比質問を生成する

表 3　比較ニュース検索の実験データ

ニュース番号		タイトル，比較質問
1	映像	米大統領、訪欧中の安倍首相と電話会談 2007/01/11
	対比質問	(イラク ∧ (行う ∨ 示す ∨ 表明)∧ ¬ (向ける ∨ 説明 ∨ 上げる ∨ 進める ∨ 取り組む ∨ 期待))
	類比質問	((向け ∧ (日本 ∨ 支援 ∨ 安倍 ∨ 総理)∧ ¬ イラク) ∨ (説明 ∧ (日本 ∨ 支援 ∨ 安倍 ∨ 総理)∧ ¬ イラク))
2	映像	トリノ五輪　環境への影響を考慮した水素バスが運行開始 2006/02/14
	対比質問	(エンジン ∧ (伝わる ∨ 感じる ∨ 起こす ∨ 登場 ∨ かける ∨ 駆動)∧ ¬ 開発)
	類比質問	((使う ∧ (スクーター ∨ 水素 ∨ オリンピック ∨ 環境 ∨ トリノ) ∪ (開発 ∧ (スクーター ∨ 水素 ∨ 路線 ∨ オリンピック ∨ 環境 ∨ 交通 ∨ 公社 ∨ トリノ)∧ ¬(エンジン ∧ バス))
3	テキスト	楽天・ＴＢＳ問題：「来月中に方向性」　村上Ｆとは「接触なし」－－ＴＢＳ 2005/10/18
	対比質問	(株 ∧ (浮上 ∨ 受け ∨ 語る ∨ 率い ∨ 買う ∨ 増す)∧ ¬ (保有 ∨ 入る ∨ 信託 ∨ 延長 ∨ でき)) ∪ (ＴＢＳ ∧ (浮上 ∨ 受け ∨ 語る ∨ 率い ∨ 買う ∨ 増す)∧ ¬ (保有 ∨ 入る ∨ 信託 ∨ 延長 ∨ でき))
	類比質問	((保有 ∧ (交渉 ∨ 村上 ∨ 期限 ∨ 楽天)∧ ¬ (株 ∧ ＴＢＳ)))
4	テキスト	阪神大震災：被災地に１２回目の祈りの朝 2007/01/17
	対比質問	(世代 ∧ (つづ ∨ 通じ ∨ 死ぬ ∨ 伝え ∨ 触れ ∨ 亡く ∨ 送る ∨ 傷つ ∨ 考え)∧ ¬ (亡くす ∨ ある ∨ でき)) ∪ (作文 ∧ (励ま ∨ 読み ∨ 書き ∨ つづ ∨ 通じ ∨ 伝え ∨ 楽し ∨ 締め ∨ 生か ∨ きる)∧ ¬ (亡くす ∨ ある ∨ ささげる ∨ 報告 ∨ 務める ∨ 成長))
	類比質問	((亡く ∧ (交流 ∨ 親 ∨ 震災 ∨ 災害 ∨ 精道 ∨ 謙)∧ ¬ (世代 ∧ 作文)))

2. ニュースアーカイブより，生成した質問を用いて検索結果を得る
3. 検索結果を正解集合により評価する
    - テキストニュースのみに対する評価
    - 映像ニュースのみに対する評価
    - すべてのメディアを用いたときの評価

実験結果を表 4 に示す．結果を以下にまとめた．

- ニュース番号 2 の対比質問では，生成した検索質問では解が得られなかった．これは，“エンジン” というキーワードに対して，“開発” という非常に共起しやすいキーワードが NOT 条件となったためである．また，被験者の作成した正解の数も非常に少なく，対比関係の検索結果として，解が無いことは正しい結果であるとも考えられる．
- 映像ニュースから生成した質問のほうが，テキストニュースから生成した質問より精度が高い結果となった．これは，テキストと映像という異メディアを対等に扱えていないことを示していると考えられ，アルゴリズムの改良を行う必要がある．
- F 値が，対比質問と類比質問で同等の値となった．つまり，対比と類比という異なる種類の質問を同精度で生成できていると考えられる．

質問生成は，キーワード重要度算出の結果に強く依存する．つまり，キーワードを誤判定している場合に結果が大きく変化する．そのため，個々のキーワード重要度のみではなく，キーワード間の関係を考慮した検索質問方式へ改良する必要がある．例えば，ニュース番号 1 であれば，“安倍” と “ブッシュ” を対等なキーワードとして扱い，これらのキーワードに基づいた比較可能なコンテンツの検索を行うという具合である．さらに，従来手法との比較や，大規模なデータを用いての実験も今後の課題としてあげられる．

6. ま と め

本稿ではニュース構成順序を用いた重要度付きコンテンツ構成グラフを定義し，そのグラフに基づいた比較コンテンツ検索のための質問生成の提案を行った．予備実験として映像ニュースとテキストニュースの構成の違いを確認した．評価実験として，コンテンツ構成グラフ生成の際のキーワードの重要度付けの妥当性を評価し，比較ニュースの検索精度を評価した．いずれも，小規模な範囲での実験にとどまっており，定量的な評価を行う必要がある．また，検索質問の生成アルゴリズムの改良と評価実験を行う必要がある．提案手法自体が，単一のニュース記事からの静的な質問生成となっている．そこで，ユーザの視聴コンテクストなどを用いて，動的に質問生成を行うことも考えられる．今後の課題としては，以下のことがあげられる．

- 大規模なニュースデータを用いた検索精度の評価実験
- 他のキーワード重要度算出手法との比較実験
- ユーザの視聴コンテクストを考慮したキーワード抽出
- キーワード間の関係に基づいた検索質問の生成
- 比較可能な提示インタフェースについての検討
- 対比・類比以外の比較関係についての検討

謝　　辞

本研究の一部は，平成 18 年度科研費基盤研究 (B)(2)「Web アーカイブと映像アーカイブを融合した次世代デジタル・ライブラリに関する研究」(課題番号：16300028)，平成 18 年度富士通 Shunsaku アカデミック支援プログラムによるものです．ここに記して謝意を表すものとします．

文　　献

[1] ANN NEWS: http://www.tv-asahi.co.jp/ann/news/web/.
[2] Asahara, M. and Matsumoto, Y.: Extended Models and Tools for High-performance Part-of-Speech Tagger., *Proc. of The 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 21–27 (2000).
[3] FNN-NEWS.COM: http://www.fnn-new.com/.
[4] Google news: http://news.google.com/.
[5] Henzinger, M., Chang, B.-W., Milch, B. and Brin, S.: Query-Free News Search., *Proc. of the 12th International World Wide Web Conference(WWW2003)*, pp. 1–10 (2003).
[6] Ide, I., Mo, H., Katayama, N. and Satoh, S.: Threading News Video Topics., *Proc. of Fifth ACM SIGMM Intl. Workshop on Multimedia Information Retrieval (MIR2003)* (2003).
[7] Ide, I., Sakai, S. and Tanaka, H.: Keyword Extraction from Various

Text Sources in News Video. (in Japanese), *Proc. of the 61th Symposium on Information Processing Society of Japan*, Vol. 3, pp. 99–100 (2000).

[8] Interstage Shunsaku Data Manager: http://interstage.fujitsu.com/jp/shunsaku/.

[9] Japanese Morphological Analysis System ChaSen: http://chasen.naist.jp/hiki/ChaSen/.

[10] Mckeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B. and Sigelman, S.: Tracking and summarizing News on a Daily Basis with Columbia's Newsblaster., *Proc. of the Human Language Technology Conference* (2002).

[11] MSN Mainichi Interactive: http://www.mainichi-msn.co.jp/.

[12] Nadamoto, A., Kondo, H. and Tanaka, K.: Web Carousel: Automatic Presentation and Semantic Restructuring of Web Search for Mobile Environments., *Proc. of the 12th International Conference on Database and Expert Systems Applications (DEXA 2001)*, pp. 712–722 (2001).

[13] Newsblaster: http://www1.cs.columbia.edu/nlp/newsblaster/.

[14] NewsInEssence: http://www.newsinessence.com/nie.cgi.

[15] Nippon Television Network News24: http://www.news24.jp/.

[16] Radev, D., Otterbacher, J., Winkel, A. and Blair-Goldensohn, S.: NewsInEssence: summarizing online news topics., *Communications of the ACM*, Vol. 48, pp. 95 – 98 (2005).

[17] Sankei Web: http://www.sankei.co.jp/.

[18] SlothLib: http://www.dl.kuis.kyoto-u.ac.jp/SlothLibWiki/.

[19] TBS NEWS i: http://news.tbs.co.jp/.

[20] Toda, H. and Kataoka, R.: A Search Result Clustering Method using Informatively Named Entities., *Proc. of the 7th annual ACM international workshop on Web information and data management*, pp. 81–86 (2005).

[21] Watanabe, T., Ohno, S., Ohta, M., Katayama, K. and Ishikawa, H.: A Distinction Emphasis Multi-document Fusion Technique. (in Japanese), *Proc. of 16th IEICE Data Engineering Workshbop (DEWS'05)* (2005).

[22] Yoshioka, Y., Yumoto, T. and Tanaka, K.: Utilizing Multimedia at News Archive by Extracting Focused Points from News Articles. (in Japanese), *IPSJ SIG Technical Reports*, 2005-DBS-137(II), pp. 415–420 (2005).

[23] Zhang, Y., He, S., Oyama, S., Tajima, K. and Tanaka, K.: Discovery of Semantically Related Topics for Given Time Series Data. (in Japanese), *DBSJ Letters*, Vol. 5, pp. 133–136 (2006).

[24] ウィキニュース:スタイルマニュアル: http://ja.wikinews.org/wiki/ウィキニュース:スタイルマニュアル.

[25] ニュースの分析: http://akasaka.cool.ne.jp/kakeru3/bs3.html.

[26] 馬強, 田中克己: 話題構造に基づく放送と Web コンテンツの統合のための検索機構., 情報処理学会論文誌, Vol. 45, pp. 18–36 (2004).

[27] 北山大輔, 角谷和俊: ニュース構成パターンに基づくビデオ・アーカイブコンテンツ閲覧方式., 電子情報通信学会第 17 回データ工学ワークショップ (DEWS'06) 論文集 (2006).