

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

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

北山 大輔[†] 角谷 和俊[‡]

Daisuke KITAYAMA Kazutoshi SUMIYA

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。

Video and text-news content have recently been broadcast on TV, newspapers, and the Internet. Although video content on out-of-date news is of little value for viewing, it can be considered to have value by comparing it to related content.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.

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。映像ニュースではTBS News i, FNN-NEWS.COMといった各報道局、テキストニュースではMSN Mainichi Interactive, Sankei Webといったインターネット上の各社のウェブサイトで公開されている。これらのサイトでの公開期間は1週間から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。

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。図1は本手法の概念図である。

以下、2節において研究の動機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. 本研究の動機と関連研究

2.1 予備実験

映像ニュースの構成は、対象物単位であるものと考えられる。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[3]。一方、テキストニュースの構成は、ニュースの内容全体を単位としたものと考えられる。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[4]。

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である。FNN-NEWS.COMはSankei Webに対応するなど、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するのは、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。
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。

実験の例を図2に示す。結果は以下のとおりである。

映像の内容構成が対象物単位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う部分から成り立っていると判断できる。

テキストの構成が概要、詳細、補足という単位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箇所になっており、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。

映像ニュースで時系列順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、概要の直後がまとめ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、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」を導くための順序

[†] 学生会員 兵庫県立大学大学院環境人間学研究所博士後期課程 ne07r001@stshse.u-hyogo.ac.jp

[‡] 正会員 兵庫県立大学環境人間学部 sumiya@shse.u-hyogo.ac.jp

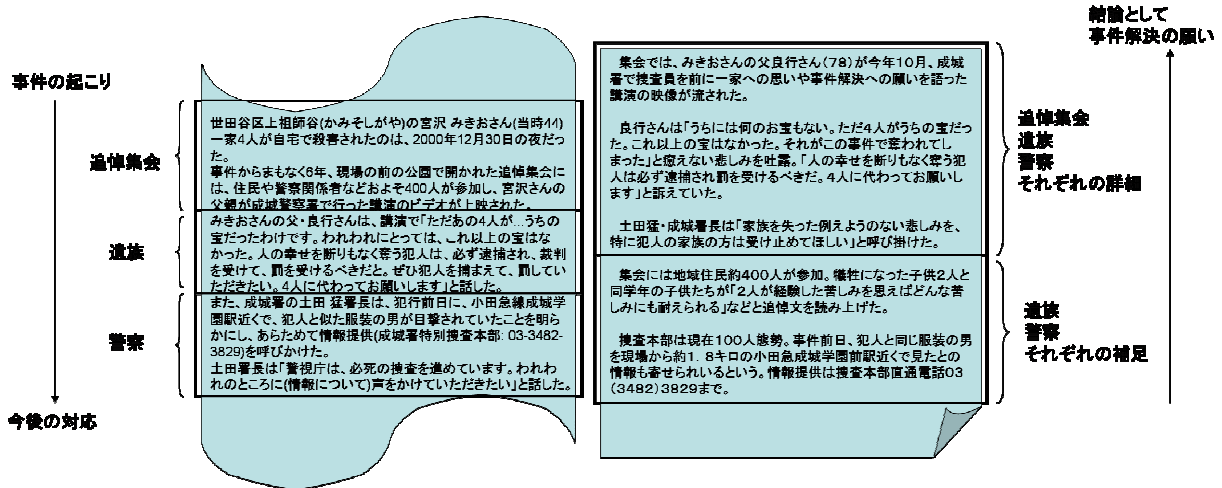


図 2 予備実験：世田谷一家殺害事件追悼集会
Fig. 2. Preliminary experiment: Setagaya family's murder memorial service

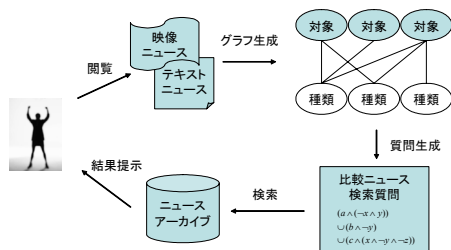


図 1 システム概要図

Fig. 1. Concept underlying comparison-news retrieval

であると判断できる。これらより、以下の4つを確認した。

1. 映像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。
2. テキストの構成単位は概要で述べられた内容であり、ニュース中の対象物が何度も述べられる。
3. 映像は時系列順に構成され、終わりでまとめを述べる。
4. テキストは結論順に構成され、はじめにまとめを述べる。

これらの特性を用いてニュースの構成を抽出し、構成が部分的に異なる比較ニュース検索の質問を生成できると考えた。

2.2 関連研究

現在ニュースサイトとしてGoogleNewsやNewsInEssenceやNewsblasterがある。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統報記事の集約・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。

コンテンツ間の関係を求める研究として、灘本ら[1]の研究があげられる。灘本らは、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、キーワードで関係を導くという点で手法が異なる。

コンテンツから検索質問を生成する研究としてHenzingerら[2]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。本研究は、内容の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にしている点で異なる。

3. 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る。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ーの種類を抽出する。また本方式では、概要部分は扱わない。

比較ニュースとは、対象や種類に着目して比較できるニュースであり、前者を対比ニュース、後者を類比ニュースと呼ぶ。ニュースの対象は名詞で表現され、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」ニュースであれば、「小泉首相」という対象と、「退任する」という表現が使われる。また、「田中知事が退任した」ニュースでは、「田中知事」という対象と、「退任する」という表現が使われる。このようにニュースカテゴリーの種類が同じであれば同じ動詞が使われると考えられる。

3.2 対象キーワード重要度の算出

対象を表すキーワードは名詞によって構成される。しかし、その出現傾向はメディアによって異なると考えられる。例えば、「小泉首相が靖国参拝をした」という映像ニュースであれば、小泉首相が出現するシーンの前後で名詞「小泉首相」が頻出することが考えられる。しかし、テキストニュースであれば「小泉首相の靖国参拝」の詳細を述べる部分や、その補足を述べる部分といった展開がされ、対象を表す名詞は記事内のさまざまな箇所に出現する。このような特徴から、映像ニュースでは以下の式により重要度算出を行う。

$$obj_val = \frac{n}{dist(a_1, a_n)} \quad (1)$$

式中の a_n は音声テキストで 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、単語の出現区間における出現密度を算出する。

テキストニュースでは以下の式により重要度算出を行う。

$$obj_val = \min \left(\sum_{i=1}^n dist(s_1, a_i), \sum_{i=1}^n dist(s_2, a_i), \dots, \sum_{i=1}^n dist(s_j, a_i), \dots, \sum_{i=1}^n dist(s_m, a_i) \right) \quad (2)$$

式中の a_i は記事中で 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。min 関数により、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。

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。

3.3 種類キーワード重要度の算出

種類を表すキーワードは動詞によって構成される。しかし、その出現傾向はメディアによって異なると考えられる。映像ニュースの内容順序の特徴として、時系列的に“何がおきた”という過去のことを述べ、“どのようにになっている”という現在のことを述べる。つまり、終端がまとめにあたると考えられる。一方テキストニュースでは、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る。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われわれは考えた。このような特徴から、映像ニュースでは、音声テキスト中で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。

$$cat_val = \sum_{i=1}^S \left(\frac{i}{S} \times count(V_i) \right) \quad (3)$$

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 中の算出対象動詞の個数を算出する。この値は後に出現するほど大きくなる。

テキストニュースでは、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。

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

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、この値は先に出現するほど大きくなり、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。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。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の出現密度の高い範囲において出現する種類キーワードが対応し、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。

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。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。

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確認を行うことができる。

対比質問の生成の様子を図3に示す。まず、ある対象とリンクしている種類を AND 条件で接続する。その際、重要度が近い語は NOT 条件とする。次に、同じ対象と接続してい

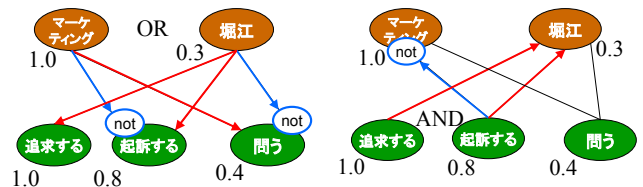


図3 対比質問の生成
Fig. 3. Comparative Query

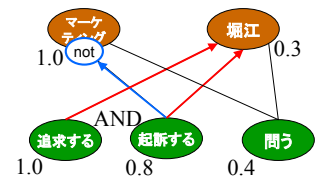


図4 類比質問の生成
Fig. 4. Analogical Query

る種類からなる質問を接続する。その際、NOT 条件のついた語、ついてない語同士を OR 条件とする。最後に、重要度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、現在見ているニュースとは異なる種類のニュースを得ることができる。

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”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。

類比質問の生成の様子を図4に示す。まず、ある種類とリンクしている対象を AND 条件で接続する。その際、近い重要度を持つ語を NOT 条件とする。次に、同じ種類と接続している対象からなる質問を接続する。その際、NOT 条件の語同士は AND 条件、それ以外は OR 条件とする。さらに、種類重要度が同じ質問を OR 条件で接続する。最後に、種類重要度が閾値以上の質問を AND 条件で接続する。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を検索結果として得ることができる。

5. 評価

5.1 プロトタイプシステム

図5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。プロトタイプでは、FNN NEWS.COM, TBS NEWS i の2種類の映像ニュースサイト、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る。ニュースアーカイブ内の検索は Interstage Shunsaku Data Manager を使い、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 を用いた茶釜による形態素解析を用い抽出した。

5.2 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。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。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利用した。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

表 1 評価実験データリスト

Table 1. News lists for experimental of retrieving comparison article

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 (イラクへ(行く)示す)表明へ(向ける)説明)上げる)進める)取り進む)期待))
	類似質問 ((向け)へ(日本)支援)安倍)総理)へ(イラク)説明)へ(日本)支援)安倍)総理)へ(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)へ(伝わる)感じる)起こす)登場)かける)駆動)へ(開発))
	類似質問 ((使う)へ(スクーター)水素)オリンピック)環境)トリノ)U(開発)へ(スクーター)水素)路線)オリンピック)環境)交通)公社)トリノ)へ(エンジン)へ(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」ーTBS 2005/10/18
	対比質問 (株)へ(浮上)受け)語る)率い)買う)増す)へ(保有)へ(入る)信託)延長)でき))
	類似質問 ((保有)へ(交渉)へ(村)上)期限)楽天)へ(株)へ(株)へ(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)へ(つづ)通じ)死ぬ)伝え)触れ)亡く)送る)傷つ)考え)へ(亡く)す)ある)でき))
	類似質問 ((亡く)へ(交流)へ(親)へ(震災)へ(災害)へ(精進)へ(謙)へ(世代)へ(作文))



図 5 プロトタイプシステム

Fig. 5. Screen image of prototype system

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の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。実験結果を表4に示す。結果を以下にまとめる。

- 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。
- 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、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、対比と類似で同等の値となった。つまり、対比と類似の質問を同精度で生成できると考えられる。

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。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。

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、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。今後の課題は、以下のとおりである。

表 2 比較実験の結果

Table 2. Experimental result of retrieving comparison article

ニュース番号	対比質問	対比質問			類似質問		
		適合率	再現率	F 値	適合率	再現率	F 値
1	テキスト	0.5	0.25	0.33	0.8	0.33	0.47
	映像	1	0.25	0.4	0.64	0.39	0.49
2	テキスト	0	0	0	0.43	0.33	0.38
	映像	0	0	0	0.44	0.71	0.54
3	テキスト	0	0	0	0.5	0.25	0.33
	映像	0.33	0.4	0.36	0	0	0
4	テキスト	0.29	0.4	0.33	0.05	0.33	0.08
	映像	0.17	0.5	0.25	0.33	0.27	0.3

- 他のキーワード重要度算出手法との比較実験
- キーワード間の関係に基づいた検索質問の生成
- 対比・類似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)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] 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).

[2] 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).

[3] ニュースの分析: <http://akasaka.cool.ne.jp/kakeru/bs3.html>.

[4] 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>.

北山 大輔 Daisuke KITAYAMA

兵庫県立大学大学院環境人間学研究所博士後期課程在学中。2007年同研究所博士前期課程修了。Web検索、映像データベースに興味をもつ。情報処理学会、日本データベース学会学生会員。

角谷 和俊 Kazutoshi SUMIYA

兵庫県立大学環境人間学部環境人間学教授。1998年神戸大学大学院自然科学研究科博士後期課程修了、博士(工学)。マルチメディアデータベース、データ放送の研究開発に従事。IEEE Computer Society, ACM, 電子情報通信学会、情報処理学会、日本データベース学会等各会員。