A Comparative News Retrieval Method using Content Elements Order

Daisuke KITAYAMA† and Kazutoshi SUMIYA†

† Graduate School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaikoen-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †ne07p001@stshse.u-hyogo.ac.jp, †sumiya@stshse.u-hyogo.ac.jp

Abstract Video and text-news content have been broadcast on TV, newspapers, and the Internet. We consider that there is a complementary relation between news articles of different media. It is possible to be promoted better understanding by comparing these articles. However, it is difficult to distinguish the complementary article from the same content article. Because the appearance structure of the keyword is only different. We propose a comparative query generation method using order of content element. It is composed of two parts. In the first part, analyzing order of news element from user’s viewing news. In the second part, generating queries using order feature of news element for retrieving comparison news.

Key words Infomation retrieval, Information integration, News archives, Query generation

1. はじめに


一般に、より深い理解を得る場合、比較するという行動を取ると考えられる。すなわち、ニュースの理解を助けるためには、ニュース中の人物や出来事、またその動作についての比較ができるコンテンツを検索して提示することが必要であると考えられる。このようなニュースは、同一内容が含まれるにもかかわらず、コンテンツの構成が異なる記事であると考えられる。例
えば、あるニュースでは主题として言われている内容が、他の
ニュースでは補足程度の扱いになっているという具合である。
また、メディアによってニュースコンテンツの構成が異なるた
め、コンテンツがどのように構成されているかを考慮した検索
を行う必要がある。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、内容が類似しているが、その
構成が変化しているニュースコンテンツを自動的に検索する仕
組みを提案する。
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構成要素の順序特性について説明し、
4節で順序に基づく比較ニュース検索を述べる。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. 研究の概要と関連研究

2.1 研究の概要

本研究では、ニュースメディアごとのコンテンツ構成要素の順
序特性を考慮した比較ニュース検索について述べる。ニュース
中より、対象となってている物事として名詞、その動作の表現と
して動詞を抽出し、コンテンツ構成要素の順序関係を付与した
質問を生成することで、ユーザが現在閲覧しているニュースコ
ンテンツの一部の内容が含まれるが、構成が変化しているよう
なニュースを検索する。このとき、構成が変化しているような
ニュースというのは、メディアのコンテンツ構成要素の順序が
異なるニュースのことであり、

コンテンツ構成要素の順序特性とは、メディアによって異なる、
文の出現順序のことを指し、映像ニュースでは、短時間で
内容を的確に伝えるために出来事の時系列的に出現させる傾向
にある。例えば、出先が先であれば時間的に前、出先が後
であれば時間的に後という役割を与えられていると捉えるこ
とが可能です。テキストニュースでは、読み返しや解釈を
などさまざまな読み方に対応するために、出来事をニュースの
理解に必要な階に出現させる傾向がある。例えば、出先が先
であればニュースの主となる内容、出先が後であれば補足的
な内容という役割を与えられていると考えることが可能です。

我々は、コンテンツ構成要素の順序特性を利用することで、
同一内容の位置づけが異なるような比較ニュース検索をするこ
とが可能であると考えた。具体的には、記事から抽出したキ
ーワードの順序関係が異なるような、異メディアの記事の検索を
行うことで、抽出したキーワードの記事内での位置づけが異
なる記事を得ることである。このようなクロスメディアの比較
ニュースを得るために、順序関係を扱える質問を定義し、一つ
の記事から自動的に順序関係を考慮した質問を生成し、検索す
る手法を提案する。この時、対象に着目した比較ニュースと、
動的に着目した比較ニュースの2つの視点の検索質問を生成
することで、比較ニュース検索を実現する。図1は本手法の概
念図である。本手法の手順は以下のとおりである。

コンテンツ構成要素の順序関係の判定 ニュース記事内のキ
ーワードの順序関係に、複数のキーワードが複数個出現するため、

一部の順序関係が定まらない、半順序関係であると考えられ
る。本手法では、構成要素である文単位で順序関係を構成され
る、2キーワード間の順序関係を判定する。判定を行う基準としては、
片方向の順序関係の出現頻度を用い、優位に出現する順序関係
を2キーワード間の順序として判定を行う。その結果、a < b, a ≠ b と
いう。 "a が b よりも先に"、"a と b の順序は同
値"、"a より b の順序が後"の関係を判定する。

順序付質問による比較ニュース検索 クロスメディア比較質問
では、通常の検索質問で用いる論理式に加え、順序関係を満た
すものを条件として付与した質問の生成を行う。さらに、対象
としての名詞を中心に比較を行う質問と、対象の起こした動作
である動きを中心に比較を行う質問の2種類を生成することで、
2つの観点から同一の出来事に関する比較を可能とする質問の
生成を行う。質問の生成は、抽出されたキーワード集合に、順
序関係を付与した状態で、論理式を生成することで行う。

このことにより、ユーザはニュースコンテンツを閲覧するた
め、自動的にニュース内容の位置づけが異なる異メディアコ
ンテンツを得ることが可能となる。

2.2 関連研究

現在提供されているニュースサイトとして McKeown [8]
GoogleNews があげられる。これらのニュースサイトは主とし
て、そのトピックを関連的に理解するための日報記事の集約・要
約が目的である。また、ニュースストリーミングを期待的に見ること
る。これらの研究は、あるニュースを多面的に見ることという点で
本研究と類似している。しかし本手法は、メディアによる構成
の差を比較するために多面的に見ることであり目的が異なる。

キーワードの出現順序に着目して検索を行う研究として、
大島 [23] や山本ら [22] の研究があげられる。大島らは「～
～～という」といった並列助詞に着目し、検索エンジンのインデク
ス中から並列助詞の前後に出現する語句を抽出することで、同
位語のクイックマイング検索を提案している。山本らは、与え
られた命題の一部に対しての真偽を評価するために命題を分割
し、一部が真で、かつ残りの命題の順序で整合した文章を
Web 中から検索することで、命題の真偽情報を提示する「ほん
と？サーチ」というシステムを開発している。これらの研究で
は、特定のキーワード出現順序を用い、その順序に合致するか
で関係を判定しているのに対し、本研究では、動的に順序関係
を抽出することで、記事の比較に適切な順序を自動的に生成した上で検索を行う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[4]や馬ら[18]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、その内容に基づいた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. コンテンツ構成要素の順序特性と比較関係

3.1 コンテンツ構成要素の順序特性

我々は、同じ内容のニュースコンテンツでも、メディアの違いによって構成に差が現れると考えた。映像ニュースでは、時間軸に沿うため、コンテンツの内容は一方的に伝達されるのかに対し、テキストニュースでは、文章を読むため、読み返しや斜め読みなど多様な伝達に対応する必要がある。このため、この構造の差は、コンテンツを構成する要素である文章の順序に現れていると考えられる。

映像ニュースの内容の順序は、短時間で内容を的確に伝えるために時間的に並びやすく、「発生（Lead-in）、現状（Body）、まとめ（Standapper）」という順序で述べられることが多い[17]。すなわち、より前に出現する文は、過去の出来事について述べている可能性が高く、逆により後に出現する文は、今後の方向性やまとめなおすといったニュース全体の総めくりを述べている可能性が高い。一方、テキストニュースの内容の順序は一般に、「概要、詳細、補足」のように、ニュースを理解するために必要な事象順（逆パラミッド型）に述べられていることが多い[16]。すなわち、より前に出る文によって、ニュースの主となる内容を述べている可能性が高く、より後に出現する文によって、ニュースの補足的な内容について述べている可能性が高い。図2にコンテンツ構成要素の順序特性の例を示す。映像では、出来事の発生から述べられ、最後に今後の方針が述べられていることがわかる。テキストでは、集中で行われた講演というニュースの主たる内容を先に述べ、文が進むにつれて、補足的な周辺内容が述べられていることがわかる。我々はこのように、メディアごとにコンテンツの内容を述べる出現順が異なることを予備実験により確認している[19]。

3.2 コンテンツ構成要素の順序特性に基づく比較関係

比較関係を抽出する目的は、ある記事と、同じ内容が含まれるが、その位置づけが異なる記事の両者を見比べることで、より深い理解を得ることにある。このような、同じ内容であるが、その位置づけが異なるようなニュースは、メディアが異なる場合に抽出可能であると考えられる。すなわち、映像ニュースで主として扱われていた内容であれば、テキストニュース中から補足として扱っている記事を抽出することで、その内容の位置づけに関しての比較を行うことが可能である。

比較を行うニュースは、比較元のニュースに対して、比較基準となるキーワードの順序が相対的に変化していると考えられる。このとき、相対的に順序が変化するというのは、動作集合に対する対象の出現順序、もしくは対象集合に対する動作の出現順序である。言い換えると、動作集合に対する名詞の出現順序、もしくは名詞集合に対する動作の出現順序といえる。図3に順序特性に基づく比較関係のモデルを示す。左は映像ニュース、右はテキストニュースである。図中の四角で囲まれたv1は名詞であり、丸で囲まれたv2、v3、v4は動詞であり、縦軸は出現順序である。このように、出現するキーワードが同じでも、メディアが異なり、順序関係が異なるようなニュースが比較関係にあたると考えられる。例えば、「安倍内閣の年金対策」を扱った映像ニュースであれば、「年金」というキーワードが比較的付け加えられ、メディアが異なる順序関係が異なるようなニュースが比較関係にあたると考えられる。例えば、「安倍内閣の年金対策」というキーワードが比較的付け加えられ、メディアが異なる順序関係が異なるようなニュースが比較関係にあたると考えられる。すなわち、このような動詞に対する、「年金」という名詞の出現順序の変化基準によって、比較関係を抽出することが可能である。

4. 順序付質問による比較ニュース検索

4.1 質問キーワードの重要度算出

クロスメディア比較質問の生成は、まず閲覧中の記事からキーワードの抽出を行い、次にキーワード間の順序関係の判定、最後に順序付質問を生成することで行う。ここでは、まずキーワード抽出のための重要度算出について説明を行う。

-279-
クロスメディアに対応したキーワード抽出を行うためには、互いのメディアを基準としたキーワード抽出を行う必要がある。すなわち、双方のメディアを統合した文書集合の中で特徴的に出現しているキーワードを抽出することが望ましい。このようなキーワードを抽出するために、各メディアの文書頻度を正規化したtf-idf値を用いる。

\[
 tf - idf = \frac{tf(k)}{df_k + \alpha \times df_k(k)}
\]  

(1)

式中の \( b \) は重要度を算出するキーワードであり、関数 \( tf \) によって現在閲覧中の記事に含まれる単語出現頻度、関数 \( df \) によって映像ニュースの中の文書頻度、関数 \( \alpha \) によってテキストニュース中の文書頻度を算出する。\( \alpha \) はメディアごとの文書数を正規化するための重みである。この式により、各メディアの文書頻度を正規化したtf-idf値を算出する。

算出したtf-idf値上位の名詞、及び動詞を用いて質問の生成を行う。このとき、質問生成のために2種類の集合を用いる。一つは、tf-idf値上位の名詞、及び動詞を含む文中にある出現する中でtf-idf値が上位の動詞のセットに、tf-idf値上位の動詞を含む文中においてtf-idf値が上位の名詞のセットに、tf-idf値上位の名詞を加えた集合、以降前者を対象集合、後者を対象集合を呼ぶ。

4.2 順序関係の判定

閲覧中の質問生成元となるニュース記事や、検索対象の記事中のキーワード間における順序関係を考慮するための判定方法について説明する。記事中から2つのキーワードを抽出し、2キーワード間の順序を判定することを行う。判定は以下の式を用いる。

\[
 \text{order}(A, B) = \begin{cases} 
 A < B & \text{(count}(a < b) > \text{count}(a > b)) \\
 A \sim B & \text{count}(a < b) = \text{count}(a > b) \\
 A > B & \text{count}(a > b) > \text{count}(a < b) 
\end{cases}
\]  

(2)

式中の \( A, B \) はそれぞれキーワード \( a \) の集合、 \( b \) の集合を表している。関数 \( \text{count}(a < b) \) により、記事中の個々のキーワード \( a \) の後の後に出現するキーワード \( b \) の個数を算出する。関数 \( \text{count}(a > b) \) は、逆にキーワード \( b \) の後に出現する \( a \) の個数を算出する。この時、前後関係の判定は文単位で行ない、同一文中に出現しているキーワードの前後関係は判定しない。すなわち、個々のキーワード間のある関係の総合で有意となる順序関係を、キーワード間の関係を、順序関係の判定を行なう。\( A < B \) は \( A \) と \( B \) の順序が \( A \) が先であることを表し、\( A \sim B \) は \( B \) が先であることを表す。また、\( A > B \) は順序関係が同値であることを表している。

4.3 順序付賃問

本手法では質問生成の際に、順序関係を用いるように拡張した論理式を用いる。我々はこの論理式を順序付賃問と呼ぶ。順序付賃問は、検索結果のコンテンツ内のにおける構成要素出現順序を考慮した論理式であり、以下のよう定義する。

\[
(a < b) = (a \land b) \lor \neg(a > b \lor (a \lor b))
\]  

(3)

\[
\neg(a < b) = \neg(a \land b) \lor (a > b \lor (a \lor b))
\]  

(4)

すなわち順序付賃問は、通常のAND条件にさらに順序関係を満たす条件を加えたものであり、より条件が厳しい質問となる。順序付賃問にNOT条件が付与された場合には、順序関係の条件を満たさないという条件しかなくなるため、順序関係を満たさなければ、キーワードを含んでいてもかまわないことに。つまり、通常のNOT条件より緩やかな質問といえる。言い換えると、AND条件部分を異なる3つの部分集合に分割し、扱うように拡張した式が順序付賃問といえる。

4.4 比較ニュース検索

クロスメディア比較ニュース検索では、対象比較、動作比較の2種類の比較質問を生成し、その結果集合を比較検索結果として提示する。まず、対象比較質問の生成について説明し、次に動作比較質問の生成について説明する。対象比較質問の生成は以下の手順で行う。

（1）対象集合中のtf-idf値上位の名詞（a）と対応する動詞をAND条件として論理式を生成

（2）個々の動詞をOR条件として論理式を生成

（3）個々の動詞と名詞aに関して抽出された順序を用いて順序付賃問を生成

(4) 3. で生成された論理式をNOT条件として、1.2.3.で生成された論理式をAND条件として論理式を生成

このようにして生成された質問より、比較軸として抽出した名詞（a）と対応する動詞を組み合わせ、かつトピックではあるが、名詞aの位置づけが異なるニュースを検索することが可能となる。すなわち、URLや媒体の表示順が変化したようなニュースを検索することが可能であるといえる。対象比較質問の生成手順を図4に示した。

次に、動作比較質問の生成は以下の手順で行う。

（1）動作集合中のtf-idf値上位の動詞（b）と対応する名詞をAND条件とした論理式を生成

（2）個々の名詞をOR条件として論理式を生成

（3）個々の名詞と動詞bに関して抽出された順序を用いて順序付賃問を生成

（4）3. で生成された論理式をNOT条件として、1.2.3.で生成された論理式をAND条件として論理式を生成

動作比較質問により、比較軸として抽出した動詞bとそれに対応する動作主体である名詞を含み、かつトピックではあるが、動作bの位置づけが異なるニュースを検索することが可能と
なる。すなわち、動作主体となった人物や出来事の動作に対する位置づけが変化したようなニュースを検索することが可能である。

5. プロトタイプシステム

5.1 システム構成

プロトタイプシステムは大きく分けて、比較質問生成部、順序関係判定部の2つの部分からなる（図5）。プロトタイプシステムの質問生成は、通常の論理式による検索質問の生成と、結果記事集合中の順序関係基準の生成を行うことで実現している。順序関係判定部では、入力キーワード集合の入力された記事における順序関係の判定を行っている。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8ヶ月間アーカイブしたものであり、ネットワーク上で配置している。ニュースアーカイブ内の検索はInterstage Shunsaku Data Manager[5]を用いた。検索のインデックスとしては、記事タイトル、記事中の名前、記事中の動詞、ニュースコンテンツの格納場所をXMLで記述したものを用意した。このため、名詞、動詞に分割をした質問を扱うことが可能である。順序関係判定部、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[14]を用いた茶塩[1]により形態素解析用に抽出した。この

<table>
<thead>
<tr>
<th>表1 名詞のt-f-idf値</th>
<th>表2 動詞のt-f-idf値</th>
</tr>
</thead>
<tbody>
<tr>
<td>名詞</td>
<td>動詞</td>
</tr>
<tr>
<td>総裁 2.03</td>
<td>誇張 0.99</td>
</tr>
<tr>
<td>きっかけ 1.89</td>
<td>結成 0.82</td>
</tr>
<tr>
<td>拘束 1.84</td>
<td>つく 0.81</td>
</tr>
<tr>
<td>新党 1.56</td>
<td>連携 0.76</td>
</tr>
<tr>
<td>民営 1.47</td>
<td>掲載 0.71</td>
</tr>
</tbody>
</table>

order(総裁,誇張) = 総裁・誇張 order(総裁,結成) = 総裁・結成 order(総裁,つく) = 総裁・つく order(総裁,連携) = 総裁・連携 order(誇張,結成) = 誇張・結成 order(誇張,つく) = 誇張・つく order(誇張,連携) = 誇張・連携 order(結成,新党) = 結成・新党 order(つく,きっかけ) = つく・きっかけ order(つく,誇張) = つく・誇張 order(つく,民営) = つく・民営 order(結成,民営) = 結成・民営

時、直後に「する」が出現する「サ変名詞」は動詞として扱っている。

図6にプロトタイプシステムの画面イメージを示す。ユーザは画面左上でニュースコンテンツを閲覧することが可能である。この時、システムは自動的にキーワード抽出を行い、比較質問を生成する。また、生成した質問を用いて自動的にニュースアーカイブに対検索を行い、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を可能とした。

5.2 自動質問生成の例

作成したプロトタイプを用いた実際の質問生成の例を説明する。質問生成に用いたニュースは、映像ニュースサイトFNN-NEWS.COMの2007年06月05日の「新党新党・総裁代表、与党が過半数割れした場合の自民党との連携に含む残る」というニュースである。この映像ニュースから抽出した名詞および動詞とそのt-f-idf値の一部を表1、2に示す。また、そのキーワードの出現順序および順序関係判定結果を図7に示す。この映像ニュースに含まれる文は5文であり、図7の区切りによって表している。図7中では、対象である名詞は四角、動作である動詞は丸で囲んでいる。この結果、対象比較質問のためのキーワード集合として、t-f-idf値が高い名詞である「総裁」が抽出される。次に、「総裁」と同じ文に出現する中でもっともt-f-idf値が高い動詞である「誇張」が抽出される。最後に、ニュース中の動詞においてt-f-idf値が高い「結成」、「つく」、「連携」が抽出される。図7の順序判定結果より、対象比較質問として表

図6 プロトタイプシステムの画面イメージ

図5 システム構成
3の質問が生成される。また、動作比較質問のためのキーワード集合として、tf-idf値最も高い動詞である「離党」が抽出される。次に、「離党」と同文に出現する中でもっともtf-idf値が高い名詞である「総理大臣」も抽出される。最後に、ニュース中の名詞においてtf-idf値が高い「きょうしつ」、「起き恒」「新党」「民営」と抽出される。図7の順序判定結果より、動作比較質問として表3の質問が生成される。

5.3 検索結果の例
生成した検索質問により得られた結果を表4に示す。対象比較質問と動作比較質問の2種類の質問生成を行っているが、双方の検索結果は同じ記事が含まれていた。生成される質問は異なるが、双方の質問に目的のキーワードの位置づけが異なる記事を検索するという目的であるため、適切に検索できている結果であると考えられる。両方で検索されている「国民新党：4月に初の政治資金パーティー」を例として、検索されたコンテンツを説明する、質問生成を行った記事では、「総理大臣」という人物を中心に記事が構成されていたが、検索結果の記事では、国民新党を中心に述べられており、「総理大臣」とは説明のために用いられているに過ぎない。しかし、「離党したが自民党との連携を考えている」といった内容は含まれており、この記事を用いて比較することにより、総理大臣個人の考えではなく、当政の考え方であるという可能性を読み取ることができる。

6.まとめ
本稿ではコンテンツ構成要素の順序特性を用いた比較質問の生成方式の提案を行った。またその中で、コンテンツ中に含まれるキーワードの順序関係について定義し、順序付け定義を行った。提案方式に基づき、プロトタイプシステムを構築し、質問生成、及び検索結果例を示し、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有効性の評価実験
- 順序を用いない質問生成との比較実験
- 比較可能な提示インタフェースについての検討

謝辞
本研究の一部は、平成19年度度富士通Shunusakkuアカデミック支援プログラムによるもので、ここに記して謝意を表すものとします。