社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－DBS－137 (II) (56)
2005／7／14

ニュースの視点の抽出による マルチメディアニュースアーカイブの利用

吉岡 由智† 湯本 高行† 田中 克己†

†京都大学情報学研究科社会情報学専攻 〒606-8501 京都市左京区吉田本町

E-mail: †{yoshitomo,yumoto,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 現在テレビやWebなどを通して膨大な数のニュース記事が配信されているが，配信されてきたニュース記事をそのまま閲覧，視聴するのではなく，一旦マルチメディアニュースアーカイブに保存しておき，そこでそれらの内容的な関係やそれを配信する者の視点の違いなどを抽出して示すことができれば，利用者はこの膨大な数のニュース記事をより有効に利用することができる．そこで本稿ではニュース記事の集合からニュース配信者の視点を抽出し，それらからトピック毎のニュースの全体像を構築する手法を提案する．トピックを伝える側の視点はトピックのテーマ，トピックの側面，トピックの側面を構成する要素からなっており，各ニュース記事から抽出される．トピックの全体像は視点を統合することによって構築する．全体像と視点を示すことによって利用者はトピックを多角的にとらえることができる．

キーワード 情報統合，データマイニング，ニュースアーカイブ

Utilizing Multimedia News Archive by Extracting Focused Points from News Articles

Yoshitomo YOSHIOKA† Takayuki YUMOTO† Katsumi TANAKA†

†Department of Social Informatics, Kyoto University

Yoshidahonmachi, Sakyo-ku, Kyoto-shi, 606-8501, Japan

E-mail:†{yoshitomo,yumoto,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract In present day, we can get enormous numbers of news articles on the Web and through TV. When news articles are stored in multimedia news archive, and the relationships of the articles and focused points of news distributors are extracted, user can use news data more effectively. In this study, we propose a way of extracting focused points of news distributors from news articles, and building an overall feature about each topic. The focused point consists of "a theme of topic", "the aspect of topic", and "the element of aspect". They are extracted from each news article. An overall feature is built by integrating focused points. By showing a overall feature and focused points about a topic, users can see the topic from various viewpoints.

1．はじめに

現在のインターネットの世界では，ウェブデータは膨大な量になっている．その中でもニュースデータの重要性は高く，今後もニュースデータの膨大化は続いていくと考えられるが，このような環境下では利用者が興味のあるニュースを閲覧するためには多くのデータの中から選別しなければならないという手間を要してしまう．そこで利用者が効率よくニュースデータを扱うことができるようにするためには，大量のニュースデータを整理しておくことが重要になる．そこで本稿ではこの状況を解決するために，それぞれのニュースデータをニュースの対象となっているトピックごとに分類することや，各トピックの内容ごとにニュースデータを分類することによってニュースデータを整理し，ニュースデータ内で述べられているトピックの全体像を利用者に提示する手法を考案した．

実用的な観点からニュースデータを見ると，Web上のニュースデータはアサヒ・コム[1]などの新聞社系のサイトやGoogle News[2]などのポータル系のサイトから配信されている．これらは速報性やニュースデータの数などといった点で優れており，利用者はいつでもあらゆるイベントに対する最新のニュースを閲覧することができる．しかし実際のニュースの利用度としてはWebニュースよりもテレビニュースの方が高く[3]，NTV[4]などの映像ニュースをWeb上で配信するサイトも整備されてきており，今後もWeb上のニュースデータの中で映像ニュースデータの占める割合は増加していくと考えられる．

従って，本研究においてもWebニュースとテレビニュースの双方のデータを扱うことにした．

また，Webニュースとテレビニュースの双方に欠落している機能として，保存期間が短いという点があげられる．Webニュースであれば，平均1ヶ月程度しか保存されておらず，テレビニュースに至ってはHDDレコーダで番組を録画していたとしても1週間程度しか保存できないという状況である．そこで本研究ではこの点を考慮して，Webニュースとテレビニュースのデータを保存しておくマルチメディアニュースアーカイブといった形でニュースデータを利用できる環境を想定している．これにより利用者はいつでも興味のあるイベントに対するニュースを閲覧することができると考えられる．

そしてマルチメディアニュースアーカイブに蓄積されたニュースデータを"郵政民営化"や"ワールドカップサッカー"などのようなニュースの対象となっているトピックの内容によって各ニュースデータを分類して整理する．そして利用者にトピックの全体像を提示することにより，利用者が1つのニュースデータを閲覧したときにそのデータがトピックのどの部分に対応したデータあるかという情報を取得することができると考えられる．またあるトピックに対して利用者が興味を持った部分に対応したニュースデータを閲覧することもできると考えられる．

トピックの全体像を構築する手法としては，共通のトピックごとにニュースデータを分類する．そのトピックを特徴付けるキーワードを抽出する．さらにトピックが"郵政民営化"のニュースデータ間でも伝えられている出来事は異なる場合がある．従って同一のトピックとして分類されたニュースデータから出来事を表すキーワードと，その出来事を説明するために用いられているキーワードを抽出する．またニュースデータを配信している配信者がトピックのどの出来事を伝えるかということは，そのトピックに対する配信者の見解であると考えられる．本研究ではこのような配信者がニュースの対象となっているトピックを見る視点というものを考慮し，トピックの出来事などと含めてトピックの内容であるとした．図1に本研究で行う処理の流れを示した．

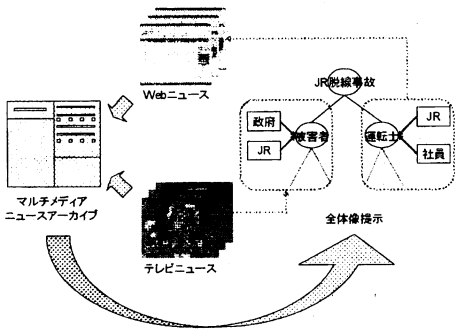

図1. 本研究の概観

以下2章では関連研究について述べ，3章ではニュースデータのトピックの内容について，4章ではトピックの全体像を構築するための手法について述べる．さらに5章では本研究で行った予備実験に対して考察し，最後に6章で本研究のまとめと将来研究について述べる．

2. 関連研究

2.1. ニュースをイベントごとに分類する研究

ニュースをイベントごとに分類することはGoogle Newsなどのように既に実用化されているものも存在するが，そのことに焦点を当てた研究も数多く存在する．TDT(Topic Detection and Tracking)[5]は，Webニュースやテレビニュース，ラジオニュースを対象にそれぞれのニュースを"郵政民営化"や"イラク戦争"といったトピックごとに分類することを目的とする研究である．またNewsBlaster[6]やNewsInEssence[7]はトピックごとに分類された各ニュース記事を融合して，要約することを目的とする研究である．しかし，本研究ではこれらの研究で行われているようなトピックごとの分類をした後に，さらに各トピックに分類されたニュース記事から出来事や配信者の視点を抽出することによって，1つのトピックの全体像を示すところが異なる．"郵政民営化"などのように世間でも注目を集めており，それに関するニュースデータも多くなるようなトピックのニュースに対しては，対象となっているトピック内でもさらに様々な出来事や見解が存在し，利用者にとってはそのトピックのニュースの全容をつかむことは困難な作業となってしまう．そのため本研究のようにあるトピックの内容の全体像を提示することによって，利用者が閲覧しているニュース記事がトピック内の記事集合の中で対応する部分がわかるということは有用である．

2.2. 話題構造の抽出

あるトピックに関するデータからトピックの階層構造を抽出する研究として小山らの手法[8]がある．この研究ではWebページ内のタイトル部に用いられている語句を主題語，本文内で用いられている語句を主題を詳細に記述するために用いられる詳細語と定義している．そしてタイトル部に含まれる主題語に対して，主題語と共起関係が高く出現頻度の高いキーワードを詳細語として抽出することによりWebページからの話題構造を抽出している．本研究においてもニュース記事がタイトル文にその記事の主題となる語が出現し，本文内に主題語を詳細に説明するための詳細語が出現するという特徴を有していることを考慮し，各ニュース記事からトピックの内容を抽出する．しかし，本研究ではさらに視点といった概念を導入することによりトピックに対する配信者の視点を抽出して分類を行うところが異なる．

3. ニュースのトピックの内容

本研究の目的はマルチメディアニュースアーカイブ内のニュースデータをトピックごとに分類して，多面的に論じられる各トピックに対する内容の比較を示したトピックの全体像を構築することにある．

本節ではトピックに対してどのような内容を抽出するかということを述べる．

(1) トピック

各ニュース記事には" 郵政民営化" などのある話題について,「民営化に反対を表明したデモが行われた.」や「小泉首相が郵政法案の修正はしないと述べた.」などの出来事に対する情報が掲載されている．このように記事内で伝えられている出来事は異なるが，記事間で共通した話題が存在する場合がある．本研究ではこのように共通した話題に関する記事集合の内容をトピックとする．

1 つのトピックにはその内容を特徴付けるテーマが存在し，そのテーマに関連した様々な出来事に対する情報が存在する．さらにテーマに関連した出来事に対してはそれを構成する要素によって分類することができる．またニュースデータを配信している配信者がトピックのどの出来事を報道するかは，トピックに対する配信者の視点を表しているとみなすことができる．以下の図.2 は" 郵政民営化" のトピックの内容を分類して全体像を示したものである.1 つのトピックに関する内容の全体像を提示することは，利用者がそのトピックの内容を知る上で有用なことであると考えられる．

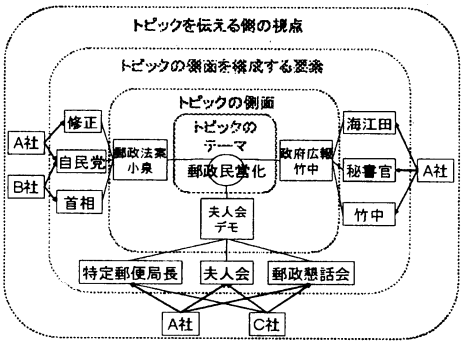

図 2. トピックの内容の分類図

本研究では図 2 のように共通のトピックの情報を掲載しているニュース記事集合から内容を抽出するために，1 つのトピックがトピックのテーマ，トピックの側面，トピックの側面を構成する要素，さらにトピックを伝える側の視点から構成されると定義した．以下それぞれについて詳細に説明する．

(2) トピックのテーマ

各ニュース記事には"郵政民営化"や"ワールドカップサッカー"などのように 1 つのテーマに関する情報が掲載されている．このように共通した話題を持つニュース記事集合を特徴付けるキーワードをトピックのテーマとする．

マルチメディアニュースアーカイブ内のニュース記事集合をトピックのテーマごとに分類し，利用者がニュース記事を閲覧したときにそのテーマを提示することによって，利用者は閲覧しているニュース記事が何について述べられているのかという情報を知ることができる．

(3) トピックの側面

それぞれのニュース記事は 1 つのテーマに関しての情報を述べているが，さらに 1 つのテーマに対しては多面的に情報が配信される．例えば郵政民営化を 1 つのトピックのテーマとしたときに「郵政民営化法案」の動向を伝えた記事や,「民営化反対で 3 千人デモ」を伝えた記事など，1 つのテーマについて全く違う出来事を伝えた記事が存在する.「郵政民営化」というトピックのテーマに対する「郵政民営化法案」や「デモ」などの出来事を 1 つのトピックに対する一部分のニュースであると考え，これらをトピックの側面とする．

各ニュース記事ではトピックの側面について詳細な情報が述べられており，トピックのテーマごとに分類された各ニュース記事からトピックの側面となるキーワードを抽出する．この際にニュース記事ではそのタイトル文に出来事を特徴付けるキーワードが出現していると考え，ニュース記事のタイトル文からトピックの側面を抽出する．そしてトピックの側面を提示することにより利用者は閲覧している記事がトピックのどの側面を伝えているものであるのかという情報を知ることができ，また違う側面からそのトピックの情報を伝えているニュース記事を閲覧することもできる．

(4) トピックの側面を構成する要素

トピックの側面が異なればニュース記事内に出現するキーワードにも違いが生じる．例えば上述した「郵政民営化法案」の側面に対しては，郵政民営化の動向を伝えるために「小泉首相」や「竹中大臣」などのキーワードが記事内に出現する．一方で「デモ」の側面に対しては，デモの内容を詳細に記述するために「特定郵便局長」や「夫人会」,「自民党郵政事業懇話会」などのキーワードが記事内に出現する．このようなトピックの側面の内容を特徴付けるキーワードをトピックの側面を構成する要素とする．

トピックの側面を構成する要素は側面を詳細に説明するために用いられるキーワードであるため，ニュース記事内の本文内に出現すると考えられる．従ってトピックの側面ごとに分類された各ニュース記事の本文からトピックの側面を構成する要素を抽出する．抽出したトピックの側面を構成する要素を利用者に提示することによって，利用者はトピックの側面がどのような要素から構成されているのかという情報がわかる．また同一のトピックの側面について伝えているニュース記事間で異なる構成要素が出現している場合に，利用者は閲覧しているニュース記事とは別の構成要素について述べているニュース記事を閲覧することもできる．

(5) トピックを伝える側の視点

各ニュース配信者が 1 つのトピックに対してのある側面について報道すれば，その配信者は報道した側面について注目しているとみなせる．このような各配信者がトピックのどの側面に注目しているかを

示すことによって，利用者は1つのトピックの全体像の中で各配信者が注目している点がわかる．例えば新聞社Aが“郵政民営化”のトピックのテーマに関して，複数個存在するトピックの側面の中で「デモ」の側面に対する報道しかしていない場合に，利用者は新聞社Aが郵政民営化に対して反対の立場を取っているのではないかと予想することができる．さらに新聞社間で報道しているトピックの側面を構成する要素に相違がある場合にも同様のことが言える．本研究ではこのようなニュース記事を配信している各配信者がトピックのどの内容を伝えているかということを，トピックを伝える側からの視点とした．

4．トピックに対する全体像の抽出

本節ではマルチメディアニュースアーカイブ内のニュースデータをトピックのテーマごとに分類し，さらに各ニュース記事からトピックの側面，トピックを構成する要素を抽出してトピックのテーマごとにニュース記事を分類する．そしてその情報に加えて新聞社ごとの見解の相違も含めた各トピックに対する全体像を提示する手法について述べる．以下の図.3に全体像を構築するまでの処理の流れを示す．

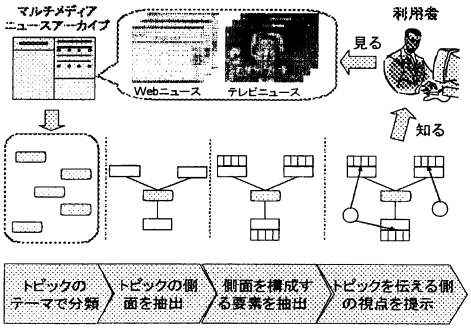

図3．トピックの全体像抽出までの処理の流れ

上図に示した流れで処理をする前段階として，マルチメディアニュースアーカイブに蓄積された各ニュース記事からその記事を特徴付けるような重要キーワード群を抽出する．一般的にニュース記事はタイトル文でその記事の対象となっているトピックの側面が述べられ，本文の冒頭段落でトピックの側面を構成する要素を含めたトピックの側面の全容が述べられる．そして第2段落以降では，トピックを構成する要素についてそれぞれ詳細に述べられているとみなすことができる．図4にニュース記事とトピックの内容の対応関係を示す．

ニュース記事のタイトル文と冒頭段落に出現するキーワードはトピックの側面，あるいはトピックの側面を構成する要素を端的に述べるために用いられている．従ってタイトル文と冒頭段落に出現するキーワードは，全てその記事において重要なキーワード

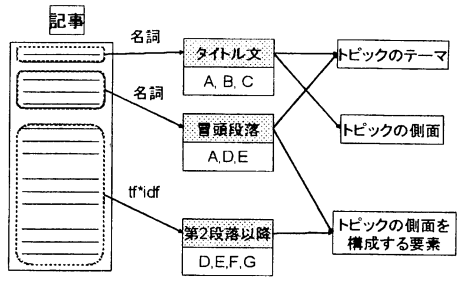

図4.ニュース記事とトピックの内容の対応

であると考えられる．よってタイトル文に含まれる名詞を抽出し，それらを重要キーワードとする．

第2段落以降からの重要キーワードの抽出には *tf*idf* 法を用いる．*tf*idf* 法はデータに含まれるキーワードの出現状況を計算に用いる手法であり，次の2つの要素を用いる．

- Term Frequency

データ中に多く含まれるキーワードはそのデータを良く特徴付ける．式は次の様に定義する．

$$tf(t,d) = freq(t,d)$$

$freq(t,d)$ はキーワード t のデータ d における出現頻度である．

- Inverse Document Frequency

キーワードは出現するデータの数が少ないほどその出現元のデータを特徴付ける．キーワード t の重要度を次の式で定義する．

$$idf(t) = \log\left(\frac{D}{freq(t,D)}\right) + 1$$

D は全データの集合を表す．

これらの2つの値を利用することでデータ d におけるキーワード t の重要度 $Weight(t,d)$ を次のように計算することができる

$$Weight(t,d) = tf(t,d) * idf(t) \quad <式1>$$

各詳細記事を *tf*idf* 法を用いてその記事に出現するキーワードに重みをつける．記事 d 中のキーワード t の重み $Weight(t,d)$ は<式 1>の値を用いる．$Weight(t,d)$ が閾値を超えるキーワードをその記事を特徴付けるキーワードとして抽出する．

4．1．記事の全体集合に対する処理

マルチメディアニュースアーカイブに蓄積したニュース記事集合からニュースの対象となっているトピックのテーマを決定し，トピックのテーマごとにニュース記事を分類する．

まず同じトピックのテーマを持つと予想されるニュース記事をまとめてトピックのテーマごとにニュース記事集合を分類する．ニュース記事間でトピックのテーマが同一になるかどうかは，ニュース記事

から抽出した重要キーワード間の類似度を求め，類似度が閾値を超えるかどうかで判断する．

トピックのテーマとなるキーワードがニュース記事中に出現する場合，その記事のタイトル文か冒頭段落に出現すると考えられる．従ってニュース記事のタイトル文と冒頭段落から重要キーワードを抽出する．そしてニュース記事 a_i におけるタイトル文と冒頭段落中の重要キーワード集合を $T(a_j)=\{t_1,t_2,...,t_m\}$ として，マルチメディアニュースアーカイブ内の記事 $A=\{a_1,a_2,...,a_n\}$ に対してキーワード間の類似度をコサイン類似度によって求め，類似度が閾値を越える記事を同じトピックのテーマの記事であるとして分類していく．

次に分類されたニュース記事中の重要キーワードの中からトピックのテーマとなるキーワードを抽出する．テーマとなるキーワードは分類されたニュース記事中のキーワードで多くのニュース記事中に出現するようなキーワードであると考えられる．

生成された1つの記事集合Cに対して，Cに含まれる記事中に出現する全ての重要キーワード集合を $T'(C)=\{t'_1,t'_2,...,t'_l\}$ として，$T'(C)$の内のキーワード t'_k が出現するニュース記事の数を $DF=freq(t'_k,C)$ として表す．そして DF が最大となるキーワードをトピックのテーマとする．

4．2．個別のニュース記事に対する処理

4．2．1．トピックの側面

トピックの側面はニュース記事内で伝えられているあるトピックの出来事を表す．トピックの側面を表すキーワードはニュース記事中でタイトル文に必ず出現すると考えられる．さらにタイトル文中のキーワードは全て重要であると考え，本研究ではニュース記事のトピックの側面をタイトル文に出現するキーワード集合からトピックのテーマとなるキーワードを除いたものとする．

ニュース記事 a_i のトピックのテーマを $Theme(a_i)$，タイトル文に出現するキーワード集合を $Title(a_i)=\{t_{t1},t_{t2},...,t_{tn}\}$ として，a_i のトピックの側面 S_i を以下の式で表現する．

$$S_i = Title(a_i) - Theme(a_i)$$

4．2．2．トピックの側面を構成する要素

トピックの側面を構成する要素はトピックの側面を説明するために用いられているキーワードである．そのため冒頭段落に出現すると考えられる．そしてその際にはトピックの側面を表すキーワードであるタイトル文のキーワードと同一の文内に出現すると考えられる．また第2段落以降では1つのトピックの側面を構成する要素に対する詳細な説明が述べられているとみなすことができる．従ってニュース記事のトピックの側面を構成する要素は冒頭段落に出現し，かつ第2段落以降にも出現するキーワードの内で，タイトル文のキーワードと同一の文内で用いられているキーワードであるといえる．

ニュース記事 a_i に出現するキーワードの中で冒頭段落かつ第2段落以降に出現するキーワード集合を $T_{element}(a_i)=\{t_{e1},t_{e2},...,t_{en}\}$ とする．そして $T_{element}$ の各キーワードとタイトル文のキーワードが同一の文内で使われている数を算出する．その値が閾値を超えるキーワード $S_{element}(a_i)=\{t'_{e1},t'_{e2},...,t'_{en}\}$ を構成する要素とする．

4．3．全体像の構築

各トピックに対する全体像の構築は各ニュース記事から抽出したトピックの側面に注目して，トピックの側面を統合することで行う．

$Theme(a_i)$ を持つ他のニュース記事からも同様にトピックの側面を抽出する．そしてトピックの側面間の類似度を求めて，類似度が閾値を超えるニュース記事を同じトピックの側面を持つ記事であるとして分類していく．その際に記事間のトピックの側面に異なるキーワードがある場合は，どちらのキーワードもトピックの側面を表すキーワードであるとする．例えば類似度が閾値を超えるニュース記事として a_1,a_2,a_3 が抽出された場合に，この3つの記事のトピックの側面がそれぞれ $S_1=\{t_1,t_2,t_3\}$，$S_2=\{t_1,t_2,t_4\}$，$S_3=\{t_1,t_2,t_5\}$ であったとすると，統合されたトピックの側面Sは以下のようになる．

$$S=S_1\cup S_2\cup S_3=\{t_1,t_2,t_3,t_4,t_5\}$$

4．4．トピックを伝える側の視点

前節までの処理によってニュースデータをトピックごとに分類できたことになる．本節では各ニュース記事を配信している配信者がトピックのどの内容を伝えているかということを提示するための手法について述べる．

各ニュース配信者についてトピックごとに分類されたニュース記事集合から抽出したトピックの側面やトピックの側面を構成する要素に対応するニュース記事があれば，その情報を提示する．これによって利用者は興味を持ったトピックに対して，各ニュース配信者にどのような内容が含まれているのかということを知ることができる．またマルチメディアニュースアーカイブに蓄積されたニュースデータの中で，それぞれのトピックの側面や側面を構成する要素に含まれるニュース記事の数を示す．これによって利用者はあるトピックに対して，配信者間で注目を集めている側面や側面を構成する要素を知ることができる．

5．予備実験

本研究では4章で述べたアルゴリズムによって，あるトピックのニュースデータから適切にトピックの側面を抽出できているということと，ニュース配信者間で同じトピックに対して伝える情報に相違があるということに対する実例を示す．2005年6月14~16日の3日間に配信されたアサヒ・コムとYOMIURI ON-LINE[9]のニュース記事のうち，両サイトで政治のカテゴリに分類された計30個のニ

ュース記事から"郵政民営化"をトピックのテーマとするニュース記事11個を抽出し,各記事のトピックの側面を比較した．この際トピックのテーマは手動で決定した．以下の表に各トピックで抽出された側面と側面に含まれる記事数を示す．

テーマ	側面	数
郵政民営化	特定郵便局長 妻 反対 3千人 デモ	1
	首相 関連 法案 国会 対応 執行部 一任	1
	民主 17日 議長 不信任案 会期 延長 対抗 紙芝居 未来	4
	首相 反論 法案 通常 国会 55日 与党 申し入れ	
	郵便局 貯金 保険 竹中 担当相 明示	1
	小泉 首相 後継 改革 路線 軌道 反対派 けん制	1
	労組 法案 廃案 集会	1
	政府 広報 契約 竹中 秘書官 関与 疑い 民主 指摘	1
	労使 交渉 労組 委員会	1

表 1.テーマ"郵政民営化"からの側面抽出

表 1 より"郵政民営化"のトピックのテーマに関して，複数の側面が抽出できている．また同一の側面に対して複数の記事が分類されている場合もあり,1つのトピックに対するニュースデータを整理できている．以下の図に表 1 で 4 件のニュース記事が分類された側面と各記事の側面を構成する要素と配信者の対応関係を示す．

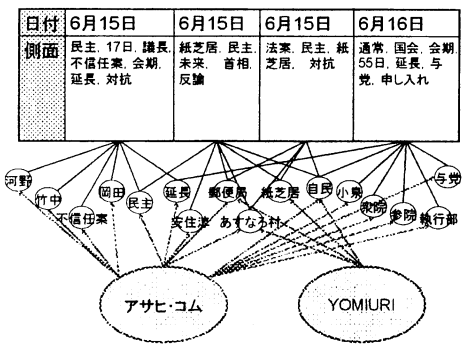

図 5. 1つの側面でのニュース記事の比較

実際に"郵政民営化"において 4 つの記事が分類された側面では，15 日のアサヒの記事が 2 件，YOMIURI の記事が 1 件，16 日のアサヒの記事が 1 件であった．それぞれの記事の内容は異なっているが，これらの記事は国会の会期延長や民主党が郵政民営化に反対の姿勢を示しているという点で同じ側面として検出されている．このように郵政民営化という共通点があるものの日付や内容が異なるデータが同じ側面として分類されるということは，ニュースデータを整理する上で意味のあることである．

6. まとめと今後の課題

本稿ではニュースデータを整理するための手法として，ニュースデータからトピックのテーマ，トピックの側面，トピックの側面を構成する要素，トピックを伝える側の視点を定義し，それらを抽出するための手法と抽出することによってトピックの全体像を提示する手法について述べた．予備実験としてトピックの側面を抽出した実例を示したが，扱ったデータ量が少ないため今後は十分なデータ量から提案した手法が有効であるかを検証する必要がある．またトピックのテーマ、トピックの側面を構成する要素を抽出できるかという実験も行う予定である．

さらに本稿で示した実例は 3 日間という短期間での Web ニュースに限定したものであった．今後はテレビニュースも含めた実験と，長期間に渡る実験を行い，マルチメディアニュースアーカイブを有効に利用できるかどうかの検証も行う予定である．

謝 辞

本研究の一部は，平成 17 年度文部科学省科学技術振興費「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」および，平成 17 年度文部科学省科学研究費特定領域研究(2)「Web の意味構造に基づく新しい Web 検索 サービス 方式に関する研究」および，２１世紀 COE プログラム知識社会基盤構築のための情報学拠点形成：「Web と放送の統合」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] http://www.asahi.com
[2] http://news.google.co.jp
[3] インターネット白書 2004 ニュースを手に入れるために利用頻度の高いメディア
[4] http://www.nnn24.com/
[5] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang: "Topic Detection and Tracking Pilot Study Final Report," Proceedings of the Broadcast News Transcription and Understanding Workshop1998
[6] Kathleen McKeown, Regina Barzilay, John Chen,David Elson, David Evans, Judith Klavans, AniNenkova, Barry Schiffman, and Sergey Sigelman.Columbia's newsblaster: New features and future directions. In *Human Language Technology Confer-nece of the North American Chapter of the Asso-ciation for Computational Linguistics(HLT-NAACL)2003 Demonstrations*, pp. 15–16, May-June 2003.
[7] Dragomir R. Radev, Sasha Blair-Goldensohn, ZhuZhang, and Revathi Sundara Raghavan.Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference*,San Diego, CA 2001
[8] Satoshi Oyama, Katsumi Tanaka, ``Query Modification by Discovering Topics from Web Page Structures,'' *Proceedings of the Sixth Asia Pacific Web Conference (APWEB'04)*, Lecture Notes in Computer Science, Vol.3007, pp.553-564, 2004.
[9] http://www.yomiuri.co.jp