T-CNB：時間を考慮した文脈に基づくニュースブラウザの提案

灘本　明代†　　田中　克己†,††

† 独立行政法人通信総合研究所　けいはんな情報通信融合研究センター
〒 619-0289 京都府相楽郡精華町光台 3-5
†† 京都大学大学院　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: †nadamoto@crl.go.jp, ††ktanaka@i.kyoto-u.ac.jp

あらまし　本論文では，新しいニュースブラウザである Time-based Contextualized-News Browser(T-CNB) を提案する．T-CNB はユーザが指定した Web 上のニュースページと関連する過去のニュースページを同じサイトから抽出し，ユーザが指定したニュースページと同時に一連のシリーズのように自動で提示する．関連する過去のニュースページの抽出にはトピックグラフを用いる．トピックグラフは主題語と内容語からなるトピックストラクチャーより作成され，メイントピックとサブトピックからなる．このメイントピックとサブトピックの類似・相違関係より関連する過去のニュースページ群を抽出する．また，T-CNB は抽出された関連ページを時間軸にそって，ひとつのシリーズのように自動提示する．T-CNB の特徴は（１）トピックグラフを用いた過去の関連ニュースページの抽出,（２）抽出された過去のニュースページ群の自動同時提示である．T-CNB を使用することにより，ユーザは閲覧したいニュースページを指定するだけで，同時に自動でそのページと関連する過去のニュースを取得することが可能となる．
キーワード　Web ブラウザ，トピックグラフ, コンテクスチュアル・ページ，　類似発見，相違発見

T-CNB: Time-based Contextualized-News Browser Based on Topic Graph

Akiyo NADAMOTO† and Katsumi TANAKA†,††

† Keihanna Human Info-Communication Research Center, Communications Research Laboratory
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
†† Department of Social Informatics,Graduate School of Informatics,Kyoto University
Yoshida Honmachi, Sakyo, Kyoto 606-8501,Japan
E-mail: †nadamoto@crl.go.jp, ††ktanaka@i.kyoto-u.ac.jp

Abstract　We propose a new way of browsing contextualized-news articles. Our prototype browser system is called a *Time-based Contextualized-News Browser* (T-CNB). The T-CNB presents user-specified news pages and related pages concurrently and automatically. It extracts the past related pages from a user-specified news articles on the web. The related pages outline the progress of user-specified news articles. We focus on the structure of the context of the page, using topic graphs to extract past related pages. A topic graph consists of a topic structure based on subject and content-describing terms. The T-CNB concurrently and automatically presents a series of related pages for one news source while browsing the user-specified page. Characteristics of the T-CNB include (1) automatically finds contextual pages from past news articles on the web based on topic graphs, (2) provides automatic and concurrent presentation of contextual pages according to time. Using the T-CNB, a user only needs to specify one news article on the web. The user then automatically receives past related news articles, which provide a wider understanding of the topic. The T-CNB automatically generates and presents contextualized news articles.
Key words　Contextualized news articles, Web browser, topic graph, similarity-detection, difference-detection

1. はじめに

現在，膨大な量のニュースが Web 上で報道され，時々刻々と更新されている．インターネットの急速なる発展に伴い，報道する側にとっても，これまでより容易に早急にニュースを取材することが可能となるとともに，これらニュースを Web 上で公開することにより，リアルタイムで人々に提示することが可能となっている．このように今日，ニュースはいつでもどこでも

取材され，人々に提示されている．現在報道されているニュースはめまぐるしく変化しているため，これらを常に注意してみていなければ，そのニュースの経緯がわからない場合が多くある．ましてや，忙しかったり，海外に行っていたりと数日間ニュースを見ていない場合，突然現在報道されているニュースを見てもその経緯がわからないため，何を報道しているのか理解できない場合がある．このような場合，我々は過去のニュースを検索し，その結果のページを順次クリックして読まなければならない．短期間に起こったニュースならばよいが，戦争のように長期的に報道されているニュースの経緯を把握するには，膨大な数のページを取得し，読まなければならない．このように，ひとつのニュースの経緯を把握するには，現在の Web ブラウザは利用者に多くの操作を要求している．CNN [1] や Yahoo ニュース [2] では関連ニュースがそのページに掲載されている場合があるが，これら関連ページがすべてのニュース項目についてあるわけではない．たとえ関連ページが掲載されていたとしても，これらのページを順次クリックして読まなければならない．現在の Web ブラウザでは，ひとつのニュースの経過を容易に，同時に取得することは不可能であり，ユーザにとっての負担が大きい．そこで我々は，ユーザがひとつのニュースソースを閲覧しているときに，そのニュースソースに関連する過去のニュースを自動で抽出し，閲覧しているニュースソースと同時に自動で提示してくれるシステムがあると便利であると考えた．本論文では，時間を考慮した文脈に基づくニュースブラウザである T-CNB(Time-based Contextualized-News Browser) を提案する．T-CNB によれば，ユーザが指定した Web 上のニュースページと関連する過去のニュースページを同じサイトより抽出し，ユーザが指定したニュースページと同時に一連のシリーズのように自動で提示する．

ほとんどのニュースは時系列データであるが，これらのコンテンツは単純にひとつの時間軸に沿って直列化されたコンテンツではなく，複数の要因が重なり合ってひとつのストーリを形成している．例えば，2003 年 9 月に起こったカリフォルニアの山火事のニュースは，最初は一箇所の火事だけではなかったため，異なるニュースとして扱われていた．しかしながら，時間がたつにつれ，これらの山火事はひとつの大きな山火事のニュースとして取り扱われた．また，時間がたつにつれ，ニュースは山火事の現状だけでなく，被害者の状況や消防士の状況，政府の対応など種々なニュースソースに分かれている．このように，ひとつのテーマのニュースは時間とともに様々な視点や立場からの報道により形成されている．そこで我々は，過去のニュースの背景の関連性を抽出することにより，ユーザが現在閲覧しているニュースの経緯が抽出できると考えた．ここでいうニュースの背景の関連性とは，ニュース間の文脈の関連性であると考え，ページ間のテーマの関連のみならず，そのテーマの背景となるサブテーマの関連性をも求める．本論文では，主題語と内容語からなるトピック・ストラクチャーから構成されるトピック・グラフを用いて，ニュース間の文脈の関連を持つページを抽出する．この抽出されたページをコンテクスチュアル・ページと呼ぶ．

この抽出された複数のコンテクスチュアル・ページを，これまでの Web ブラウザのように提示したのではユーザにとってニュースの経緯を容易に取得できるとは限らない．そこで，T-CNB ではこの抽出されたコンテクスチュアル・ページを自動で提示することとともに，音声読み上げを用いて受動的に取得できるようにする．T-CNB の画面イメージを図 1 に示す．左ウィンドウはユーザが指定したニュースページを示し，右ウィンドウはそのニュースページと関連するコンテクスチュアル・ページを示す．T-CNB はコンテクスチュアル・ページはページのタイトル，画像，及びページの内容の部分からなるページ・コンポーネントを提示する．このように，ユーザは順次閲覧したいページをクリックするだけで，T-CNB は現在ユーザが閲覧しているページの経緯を示す過去のニュース群を抽出し，自動で紙芝居のように複数のページ・コンポーネント群を提示しながら，そのページ・コンポーネントを構成するページの部分を音声読み上げを行う．

T-CNB の特徴を以下に示す．

- トピック・グラフを用いたコンテクスチュアル・ページの抽出
  - ユーザが指定したページのコンテクストを示す過去のニュースページ群の抽出
  - 抽出した Web ページ間の類似・相違によるコンテクスチュアル・ページの抽出
- 抽出したページ群の自動提示
  - ページ・コンポーネントの生成および自動提示
  - コンテクスチュアル・ページの受動的視聴
  - コンテクスチュアル・ページ群の自動提示

以下，2 章では関連研究を，3 章では T-CNB の基本コンセプトを，4 章ではコンテクスチュアル・ページの抽出方法を，5 章で実験結果について述べ，6 章でまとめと今後の課題について述べる．

2. 関 連 研 究

複数のニュース記事の類似に関する研究は多くある [7] [8] [9] が，これらの研究のほとんどは，ニュース記事の要約を目的としている．我々の研究は複数のニュース記事の類似，相違から関連するニュースページを抽出し，これらを自動的に提示するため目的が異なる．また，トピックの抽出の研究では，The Topic Detection and Tracking (TDT) project [10] がある．このプロジェクトでは多くの event detection and event tracking に関する研究が行われている [11], [12], [13]．我々の研究は event detection を行っているが，ほとんどの event detection に関する研究は自然言語処理に基づくものであり，我々の研究は主題語，内容語をひとつの構造としてみてトピックグラフに基づき event detection を行っている点が異なる．また，ニュースページのブラウジングに関する研究も多く行われている．Columbia's Newsblaster [14] [15] は event tracking を行い，複数ニュースの要約を行っている．彼らは，我々と同様に類似・相違の発見に注目しているが，これらに基づき複数ニュースの要約を行い，ひとつのウィンドウでその要約を提示している．これに対し我々

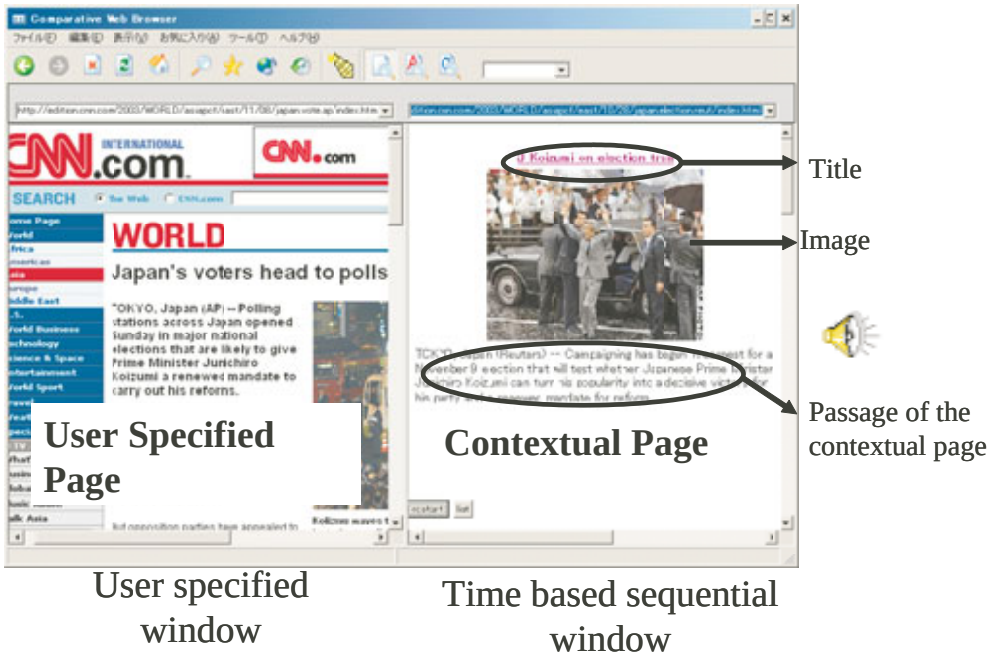

図 1　画面イメージ

Fig. 1　Picture of Display

は，類似・相違の発見を行いこれらの関係に基づき，関連するニュースを抽出し，これらをひとつのシリーズのように自動で提示している．また，Columbia's Newsblaster は自然言語処理にもとづく類似・相違発見を行っているが，我々はトピックグラフに基づき類似・相違発見を行っている点が異なる．

これまで我々は，2 つの異なる Web サイトから類似するページを発見し同時に比較提示する新しいブラウザである the Comparative Web Browser(CWB) [16] と，その多言語対応である the Bilingual Comparative Web Browser(B-CWB) [17] を提案してきた．これらで使用したトピックストラクチャーを T-CNB では使用しているが，CWB と B-CWB はトピックストラクチャーの主題語，内容語間の類似度をユークリッド距離を用いて求めているのに対し，T-CNB では，トピックストラクチャーからトピックグラフを生成し，そのページのメイントピックとサブトピックを抽出し，類似度・相違度を求めている点が異なる．また，CWB，B-CWB ではユーザの指定したページのオペレーションに対し同期するように比較するページを提示しているが，T-CNB では，ユーザの指定したページとそれに関連する複数のページを順次提示する方法が異なる．

3.　基本コンセプト

T-CNB の画面イメージを図 1 に示し，以下に基本コンセプトを述べる．

3.1　コンテクスチュアル・ページの概念

T-CNB はユーザが指定したニュースページと関連する過去のニュースページを抽出し，これらを自動的に提示するシステムである．しかしながら，関連する過去のページをすべて提示したのでは，それらのページが多すぎる場合がある．そこで，抽出した関連する過去のページから，ある程度変化があり，これらのページを時系列につなげると，ユーザの指定した現在のニュースの経緯がわかるページ群を抽出する必要がある．我々は，ニュースページはメインとなるトピックとその内容や背景を示すサブトピックから構成されていると考えこのメイントピックとサブトピックに注目する．時系列にニュースが変化する時，あるページではメイントピックであったものが，次の時刻のニュースではサブトピックになる場合がある．また，その逆も考えられる．そこで我々は，トピックグラフを用いて，ひとつのニュースソースのトピックの変化を求め比較する．ここでいう，メイントピックは複数の単語からなりそのページを顕著に示す単語群を示す．サブトピックは複数の単語からなりそのページの内容や背景を示す単語群を示す．本論文ではトピックグラフを用いてメイントピックとサブトピックを抽出する．T-CNB では比較元のページのメイントピックと類似し且つ，比較元のサブトピックと相違するトピックグラフを持つページをコンテクスチュアル・ページとする．つまり T-CNB はページ同士のメイントピック，サブトピックの類似度，相違度から各々のページ間の話題の変異を求め，どのページをコンテクスチュアル・ページとするかを決定する．

3.2　コンテクスチュアル・ページの提示方法

T-CNB はユーザが指定したページを閲覧しているときに，同時にユーザに関連する複数のページを提示する．このとき，従来の Web ブラウザのように複数のページを提示したのでは，ユーザは複数の操作を同時に行い，またそれらのページも読まなければならい．これは膨大な時間がかかるとともに，一度に複数のページの意味や経緯を理解することは困難でありユーザにとって負担が大きい．我々は，ユーザが指定したページを閲覧しながら，容易にわかりやすく関連する複数のページの情報を取得するには，ページのすべてを提示するのではなくページ

の部分を提示するとともに，音声読み上げを用いてユーザが受動的にコンテンツを取得できるようにするのが好ましいと考えた．そこで T-CNB ではその Web ページを顕著に示す部分からなるページ・コンポーネントを音声読み上げを用いて提示し，これら複数のページ・コンポーネントをひとつのストーリのように提示する．

ページ・コンポーネント

ユーザが一目でその Web ページが何を示しているのかを理解するのは，そのページのタイトルと画像の提示が好ましい．そこで T-CNB では，ページのタイトル，画像，コンテンツの部分からなるページ・コンポーネントを作成し提示する．

T-CNB ではページのタイトルに $< title >$ タグを用いず，Web ページの構成上タイトルと予測される文字列をタイトルとする．実際には，$< Font >$ タグまたは $< H >$ タグで囲まれた単語もしくは文であり且つ文末の単語が名詞もしくは固有名詞であるものをそのページのタイトルとする．

ページ・コンポーネントではそのページの要約を読み上げる．ニュースページの場合，最初の段落がその記事の要約を示している場合がほとんどであるため，T-CNB では，ページの最初の段落の文字列を提示するとともに，この部分の音声読み上げを行う．しかしながら，実際には要約を示すだけでなく，他のページとの差分情報も示した方が好いため，どの部分を音声読み上げするかは今後の課題である．

画像は，ユーザが一目でみてそのページが何を報道しているの把握できる要因のひとつである．そのため，ページ・コンポーネントでは画像を大きく示して提示する．図 2(a) に示すように，1 つのページに複数の画像がある場合，上記で述べたテキストを音声読み上げしている間に順次提示する．

このように，T-CNB は抽出した関連ページの部分を示すページ・コンポーネントを生成し，これらを提示する．

複数ページ・コンポーネントの受動的提示

1 章で述べたように，ニュースソースは時系列的に関連する複数のニュースがひとつのストーリーのようになっている．システムが抽出した各々のコンテクスチュアル・ページはこのシリーズの中に点在する要素である．つまりは，ユーザはこれら複数のコンテクスチュアル・ページを順次閲覧することにより現在自分が指定したニュースの経緯が把握できる．そこで，T-CNB ではこれら複数のコンテクスチュアル・ページから構成されている複数のページ・コンポーネントをその記事が報道された時間軸に沿って自動て提示する．この時，図 2(b) に示すように，T-CNB は音声読み上げが終了したページ・コンポーネントをフェイドアウトし，次のページ・コンポーネントを提示する．フェイドアウトの効果を用いることにより，ユーザにやんわりと記事の変化を知らせることが可能となる．また，すべてのページ・コンポーネントの提示終了後，T-CNB はこれらページ・コンポーネントのリストを右ウィンドウに示す（図 2(c)）．ユーザがプレビューボタンをクリックすることにより，システムは再度ページ・コンポーネントを提示する．

3.3 プロトタイプシステム

我々は，T-CNB のプロトタイプシステムを開発した．開発言語には Microsoft C#を用いた．図 3 にプロトタイプシステムの構成を，以下にシステムの流れを示す．

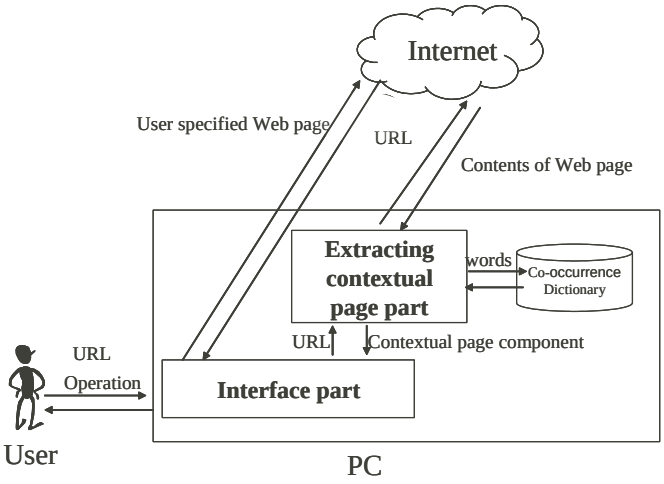

図 3　プロトタイプシステム構成図
Fig. 3　System Architecure

（1） URL の指定

ユーザは閲覧したいニュースサイトの URL と関連ページを検索したい期間を入力する．そして，システムはユーザが指定したページを取得し左ウィンドウに提示する．

（2） ユーザの指定したページから検索キーワードの抽出

システムはユーザが指定したページから主題語と内容語からなるトピックストラクチャーを抽出し，トピックグラフを生成する．このユーザが指定したページのトピックグラフから最初のコンテクスチュアル・ページを発見するための検索キーワードを抽出する．

（3） コンテクスチュアル・ページの抽出

システムは抽出した検索キーワードを用いてコンテクスチュアル・ページを検索する．ここで検索元のメイントピックと類似しており，サブトピックと異なるトピックストラクチャーを持つページがコンテクスチュアル・ページとなる．

（4） コンテクスチュアル・ページから検索キーワードの抽出

システムは (3) で抽出されたコンテクスチュアル・ページからトピックストラクチャーを抽出後トピックグラフを生成し，そして次の検索のための検索キーワードを決定する．

（5） ユーザが指定した期間の間または，類似度及び相違度がある閾値の間システムは (3) と (4) を繰り返す．

（6） ページ・コンポーネントの生成

すべてのコンテクスチュアル・ページの検索終了後，各々のページのページ・コンポーネントを作成する．

（7） ページ・コンポーネントの提示

複数のページ・コンポーネントを自動で受動的にユーザに提示する．

4. コンテクスチュアル・ページの抽出

T-CNB はトピックグラフを用いてコンテクスチュアル・ページを抽出する．トピックグラフはトピックストラクチャーから

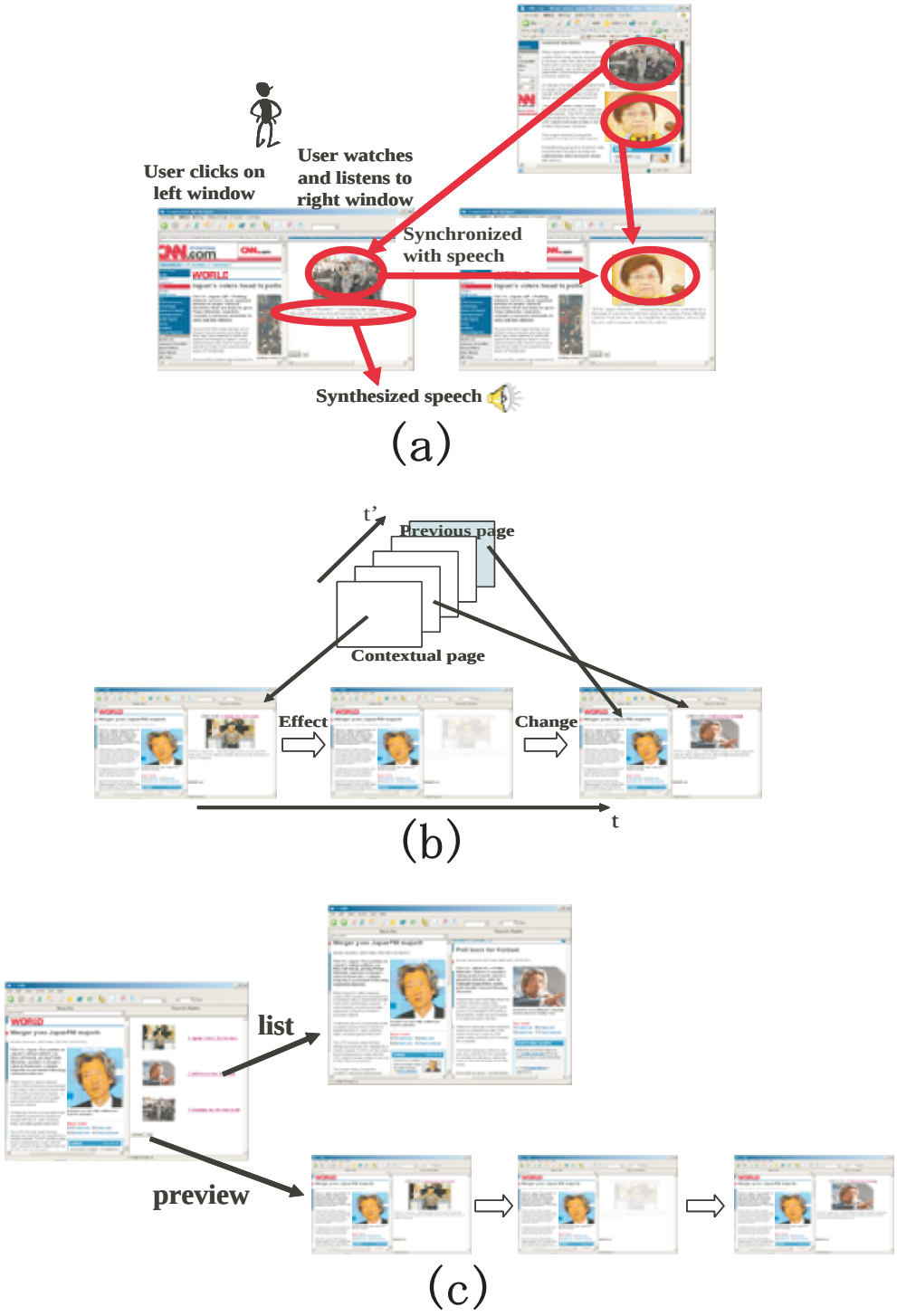

図 2　ユーザインターフェイス
Fig. 2　User Interface

なり，そのトピックストラクチャーはページの主題語，内容語からなる．

4.1　トピックストラクチャーの生成

T-CNB では，松倉ら [5] が提案した単語の共起関係を用いたトピックストラクチャーに基づいた方法を用いる．ページ P におけるトピックストラクチャー TP はトピック $t_i, i \in \{1, \ldots, n\}$ からなり，t_i は主題語 s_i と内容語の集合 C_i の 2 つの組からなる．また，C_i は複数の内容語 $c_{im}, m \in \{1, \ldots, k\}$ で構成される．すなわち，TP は以下のとおりである．

$$TP = \{t_1, ..., t_i, ..., t_n\}$$
$$t_i = (s_i, C_i)$$
$$C_i = (c_{i1}, .., c_{im})$$

主題語の抽出

松倉らは，主題語をページにおいて単語の密度が高い単語としている．T-CNB ではページ単位での類似関係を求めるため，単語の出現頻度を用いて抽出する．また，対象となる単語は名詞のみとする．すなわち，主題語の候補となる単語 t は

$tf(t) \times weight(t) > \alpha$

となる．ここで，$tf(t)$ は P における t の出現頻度を示し，$weight(t)$ は品詞による単語の重みを示し，α は閾値を示す．我々は実験により，$weight(t)$ を固有名詞は 3.0，数及び助数詞は 0.1，一般名詞は 1.0，その他の名詞は 0.9 とした．

内容語の抽出

内容語は主題語との共起度の高い単語とする．我々はあらかじめニュースにおける単語の共起辞書を作成し，この共起辞書を用いて共起度を求める．P の主題語を $\{s_1, \ldots, s_i, \ldots, s_n\}$ とすると，各々の主題語 s_i において内容語の集合である $C_i = \{c_{i1}, \ldots, c_{ij}\}$ を求める．c_{ij} は,s_i との単語の共起度がある閾値 (β) 以上の単語である．内容語も名詞のみを対象とする．

このように，ページ P における主題語と内容語を決定する．よって，P のトピックストラクチャーは $\{t_1 = (s_1, C_1), ..., t_n = (s_n, C_n)\}$ となる．

4.2 検索キーワードの抽出

トピックグラフの生成

T-CNB は上記で求めた主題語をルートとし，その子節点を内容語とするグラフを作成する．1 ページのトピックストラクチャーには複数の主題語が含まれているため，複数のグラフが生成される．そして，各々のグラフのルートが他のグラフの子節点となっている場合，これらを結合する．このようにしてトピックグラフを生成する．トピックグラフは無向グラフであり複数の連結成分を持つ．

例えば，ブッシュ大統領がカリフォルニアの山火事の視察に行ったニュースページ P の場合，トピックストラクチャーは以下のようになる．

$TP = \{t1, t2, t3\}$

$t1 = \{s_1, (c_1, c_2)\} = \{$ カリフォルニア, (山火事, 消防士)$\}$

$t2 = \{s_2, (c_3, c_4, c_5)\}$

$= \{$ 消防士, (火事, カリフォルニア, 緊急)$\}$

$t3 = \{s_3, (c_6, c_7)\} = \{$ ブッシュ, (大統領, ツアー)$\}$

このページ P のトピックグラフを図 4 に示す．この場合，”カリフォルニア”と”消防士”は主題語であり且つ各々の内容語でもあり，各々のグラフに出現している．そこで，グラフ (t1) と (t2) を結合し，ひとつの連結成分 (a) とする．すなわち，ページ P のトピックグラフは 2 つの連結成分からなる．

この時，T-CNB はニュースのトピックの移り変わりを抽出したいため，”山火事”のように内容語であり主題語でない節点が複数のグラフに出現しても結合せずに無視する．

検索キーワードの抽出

作成したトピックグラフの中で最も節点数の多い連結成分はそのページの特徴を示す単語の集合であると考え，そのページのメイントピックとする．また，メイントピック以外の連結成分をサブトピックとする．T-CNB はこのメイントピックを構成する主題語と内容語を用いて，コンテクスチュアル・ページを検索するための検索キーワードを抽出する．このように，ページ P の検索キーワード Q は以下のように決定する．

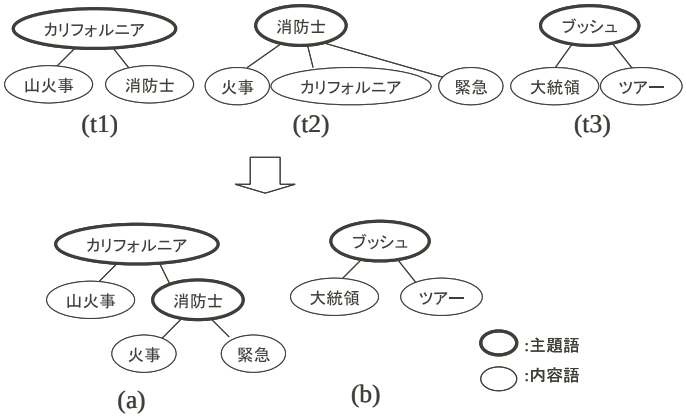

図 4　トピックグラフの例

Fig. 4　Example of Topic Graph

$$Q = (s_1 \wedge s_2 \wedge \cdots \wedge s_n) \wedge (c_1 \vee c_2 \vee \cdots \vee c_m)$$

ここで，$s_1, \cdots, s_n$，$c_1, \cdots, c_m$ はページ P における最大節点数を持つ連結成分中の主題語と内容語を示す．図 4 の場合，検索キーワードは (カリフォルニア ∧ 消防士) ∧ (山火事 ∨ 火事 ∨ 緊急) となる．

4.3 コンテクスチュアル・ページの抽出

T-CNB は検索キーワードを用いて，ユーザの指定しているニュースページと同じサイト内の過去のニュースページから関連するページを検索する．この検索されたページがコンテクスチュアル・ページの候補となる．トピックグラフはメイントピックとサブトピックから構成されている．サブトピックはそのニュースページの内容や背景を示す情報である．

T-CNB はユーザの指定したニュースの経緯をユーザが簡潔に理解できるようにすることを目的としているため，ある程度異なる情報を持つページを提示することが望ましいと考える．そこで，このトピックグラフの構造に注目し，コンテクスチュアル・ページを抽出する．

ニュースが時系列に変化していく場合にはいろいろな変化が考えられるが，大まかに以下の 2 つに分類することができる．

（1） 過去の記事では全般的に取り扱われていたものが，現在の記事ではメインテーマとなる場合

例えば，事件，事故のニュース記事によくあるように，最初は犯人がわからず事件の全容のみを記述したニュースが報道されているが，時間がたつにつれて犯人が判明し，その後その犯人の人間像なども報道されてゆく場合などである．この場合，ページ P と，ページ P' の関係は以下のようになる．

- ページ P のメイントピックと，ページ P' のメイントピック及びサブトピックの類似度が高いページ P'.
- ページ P のサブトピックと，ページ P' のメイントピック及びサブトピックの相違度が高いページ P'.

（2） 過去の記事ではメインテーマであったものが，現在の記事では全般的に取り扱われている場合

例えば，イラク戦争のように最初はイラク戦争がメインテーマであったのに対し，現在はイラク戦争の戦後の種々な処理がテーマとなり，イラク戦争そのものがメインテーマとなってい

ない場合などである．この場合，ページ P と，ページ P' の関係は以下のようになる．

- ページ P のメイントピック及びサブトピックと，ページ P' のメイントピックの類似度が高いページ P'．
- ページ P のメイントピック及びサブトピックと，ページ P' のサブトピックの相違度が高いページ P'．

本論文では，第一段階として，(1) の過去の記事では全般的に取り扱われていたものが，現在の記事ではメインテーマとなっている過去のページをコンテクスチュアル・ページとする．

類似度及び相違度は以下のように求める．

類似度

T-CNB では，検索元となるページ P のメイントピックと検索結果である過去のニュースページ P' のメイントピック及びサブトピックの類似度を求める．G をページ P のトピックグラフとし，G' をページ P' のトピックグラフとすると，

$G = G_1 \cup \cdots \cup G_i \cup \cdots \cup G_n$

and

$G' = G'_1 \cup \cdots \cup G'_j \cup \cdots \cup G'_m$

となる．ここで，$G_i (i \in \{1, \cdots, n\})$ と $G'_j (j \in \{1, \cdots, m\})$ は各々G と G' の連結成分を示す．また，G_1 と G'_1 は各々のページにおける最大の連結成分であるメイントピックを示す．G' と G の類似度を以下に示す．

$$Sim(G', G) = \frac{1}{m} \sum_{j=1}^{m} \frac{\mid v(G'_j) \cap v(G_1) \mid}{\mid v(G'_j) \cup v(G_1) \mid}$$

$v(G)$ はグラフ G の各節点を示す．つまりは，ページ P のメイントピックを構成している主題語，内容語がページ P' のトピックグラフにどれだけ含まれているかが類似度である．上記のように T-CNB はページ P とページ P' との類似度を求め，類似度 Sim が γ 以上のページ P' をコンテクスチュアル・ページの候補とする．

相違度

T-CNB では，検索元となるページ P のサブトピックと検索結果である過去のニュースページ P' のメイントピック及びサブトピックの相違度を求める．G と G' の相違度は以下の通りである．

$$Diff(G', G) = 1 - \frac{1}{(n-1)m} \sum_{i=2}^{n} \sum_{j=1}^{m} \frac{\mid v(G'_j) \cap v(G_i) \mid}{\mid v(G'_j) \cup v(G_i) \mid}$$

ページ P のサブトピックを構成している主題語，内容語がページ P' のトピックグラフにどれだけ含まれていないかが相違度である．相違度 $Diff$ が δ 以上のページ P' をコンテクスチュアル・ページの候補とする．

このようにして，類似度，相違度を求め，おのおのの値が閾値以上のページをコンテクスチュアル・ページの候補とする．この候補の中から，最もページの作成時間が検索元のページ P に近いものをコンテクスチュアル・ページと決定する．

システムは，最初はユーザの指定したページからトピックグラフを生成し，そのページに関連するコンテクスチュアル・ページを決定する．次に，その決定されたコンテクスチュアル・ページのトピックグラフを用いて，そのページに関連するコンテクスチュアル・ページを決定する．このように，T-CNB は時間をさかのぼりコンテクスチュアル・ページを抽出してゆき，コンテクスチュアル・ページの候補となるページがなくなるまで，またはユーザが指定した期間まで抽出を行う．

表 1　類似度，相違度の閾値における適合率 (precision %)

Table 1　Precision Ratio for Similarity Degree and Difference Degree (precision %)

類似度	相違度	適合率
0.3	0.3	19
	0.5	23
	0.7	24
0.5	0.3	52
	0.5	57
	0.7	42
0.7	0.3	42
	0.5	51
	0.7	18

5. 実　　験

プロトタイプシステムを用いて，以下の 2 つの実験を行った．

類似度と相違度の閾値の実験

トピックストラクチャーを求めるコンテクスチュアル・ページを抽出するための，ページ間の類似度，相違度の最適な閾値を求める実験を行った．実験は，CNN [1] と Yahoo ニュースサイト [2] から 1000 ページを対象に，類似度と相違度の閾値を変え，抽出されたコンテクスチュアル・ページの適合率を求めた．この時，トピックストラクチャーを構成する主題語の閾値（単語の出現頻度）を 6 とし，内容語の閾値（共起度）を 0.3 とした．実験結果を表 1 に示す．この実験より，類似度の閾値を 0.5，相違度の閾値を 0.5 とした時，最も適合率が高いことがわかった．

サイト毎のコンテクスチュアル・ページの適合率

我々は，CNN ニュース [1]，Yahoo ニュース [2]，USA TODAY [3]，CBS NEWS [4] の 4 つのニュースサイト合計任意の 200 ページを指定し，各々のサイトにおける抽出されたコンテクスチュアル・ページ群の適合率の実験を行った．実験結果を表 2 に示す．適合率の高いのページは，上記アメリカのニュースサイトの中で日本の選挙のニュースのように，大変特徴があり，また短期間シリーズとして報道されているニュース記事のコンテクスチュアル・ページであった．

表 2　サイトにおけるコンテクスチュアル・ページの適合率 (%)

Table 2　Precision Ratio of Contextual Pages(%)

ニュースサイト名	適合率
CNN	55
Yahoo news	54
USA TODAY [3]	51
CBS NEWS [4]	50

6. ま と め

本論文では，時間を考慮した文脈に基づくニュースブラウザである T-CNB(Time-based Contextualized-News Browser) を提案した．T-CNB は，ユーザが指定した Web 上のニュースページと関連する過去のニュースページを同じサイトより抽出し，ユーザが指定したニュースページと同時に一連のシリーズのように自動で提示する新しい Web ブラウザである．我々は，関連する過去のニュースページをコンテクスチュアル・ページと呼び，主題語と内容語からなるトピックストラクチャーにより構成されるトピックグラフを用い，そのトピックグラフの類似，相違性からコンテクスチュアル・ページを抽出した．また，その抽出したコンテクスチュアル・ページをユーザが閲覧しているページと同時に自動で受動的に提示した．

T-CNB の特徴を以下に示す．

- トピック・グラフを用いたコンテクスチュアル・ページの抽出
  - ユーザが指定したページのコンテクストを示す過去のニュースページ群の抽出
  - 抽出した Web ページ間の類似・相違によるコンテクスチュアル・ページの抽出
- 抽出したページ群の自動提示
  - ページ・コンポーネントの生成および自動提示
  - コンテクスチュアル・ページの受動的視聴
  - コンテクスチュアル・ページ群の自動提示

ユーザは閲覧したいニュースページを指定するだけで，同時に自動でそのページと関連する過去のニュースを取得することが可能となる．

今後の課題を以下に示す．

- コンテクスチュアル・ページの抽出方法の更なる検討

4 章で述べたように，種々なコンテクスチュアル・ページの抽出方法が考えられる．本論文では，その第 1 弾としての手法を提案した．そこでこの種々な抽出方法による実験を行い，最適な手法の検討を行う．

- コンテクスチュアル・ページの提示方法の更なる検討

現在の提示方法ではコンテクスチュアル・ページとユーザが指定しているページとの相違点を示していない．そこで今後，この相違点を提示する方法の検討を行う．そして，現在ページコンポーネントで音声読み上げを行う部分は，ページの第一段落である．しかしながら，コンテクスチュアル・ページ間の類似・相違点を示す文章を読み上げたほうがよりユーザにとってわかりやすいと考え，音声読み上げを行う文章の検討を行う．また，画像がないページの場合，現在の T-CNB ではタイトルと音声読み上げを行う文章のみの提示となり，一目でそのページが何を示しているのかわかりにくい．そこで，画像のないページの場合，そのページに関連する画像を他のページから抽出し提示する方法を検討する．

文 献

[1] CNN site homepage http://www.cnn.com

[2] Yahoo news site homepage http://news.yahoo.com/

[3] USA TODAY site homepage http://www.usatoday.com

[4] CBS NEWS site homepage http://www.cbsnews.com/

[5] T.Matsukura, H.Kondo, Y.Hirata, and K.Tanaka, ”Discovery of semantic relationship among web pages based on web topic structures”, Proc. of 9th IFIP 2.6 Working Conference on Database Semantics, 2001.

[6] Asahi newspaper site homepage http://www.asahi.com

[7] D.R.Radev, H.Jing, M.Budzikowska, ”Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies”, Proc. of ANLP-NAACL 2000, Seattle, pp21-29, Washington, May 2000.

[8] J.Carbonell and J.Goldstein, ”The use of MMR, diversity-based reranking for reordering documents and producing summaries”, Proc. of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp335-336, Melbourne, Australia, August 1998.

[9] J.Goldstein, M.Kantrowitz, V.Mittal and J.Carbonell, ”Summarizing Text Documents: Sentence selection and evaluation metrics”, Proc. of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, August 1999.

[10] TDT site homepage http://www.itl.nist.gov/iaui/894.01/tests/tdt/index.htm

[11] M.Spitters and W.Kraaij, ”A Language Modeling Approach to Tracking News Events,” TDT 2002 Evaluation workshop, Gaithersburg, MD, USA, 2002.

[12] J.Allan, R.Papka, V.Lavrenko,”On-line New Event Detection and Tracking”, Proc. of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.37-45, Melbourne, Australia, August 1998.

[13] Y.Yang, T.Pierce and J.Carbonell, ”Study on Retrospective and On-Line Event Detection”, Proc. of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.28-36, Melbourne, Australia, August 1998.

[14] Columbia’s NewsBlaster site homepage http://www1.cs.columbia.edu/nlp/newsblaster/

[15] K.McKeown, R.Barzilay, D.Evans, V.Hatzivassiloglou, J.Klavans, C.Sable, B.Schiffman and S.Sigelman , ”Tracking and Summarizing News on a Daily Basis with Columbia’s Newsblaster”, Proc. of the 2002 Human Language Technology Conference (HLT). San Diego, California.

[16] A.Nadamoto and K.Tanaka, ”A Comparative Web Browser (CWB) for Browsing and Comparing Web Pages”, The 12th International World Wide Web Conference (WWW2003), pp.727-235, Budapest, Hungary, May 2003

[17] Akiyo Nadamoto, Ma Qiang, and Katsumi Tanaka, ”Concurrent Browsing of Bilingual Web Sites By Content-Synchronization and Difference-Detection”, Proceedings of the 4th International Conference on Web Information Systems Engineering (WISE2003) , pp.189-199, Roma, Italy, Dec 2003.