B-CWB:類似コンテンツの視点差異情報を同時提示する多言語Webブラウザ

B-CWB: Multilingual Web Browser for Concurrent Browsing of Different Perspectives of Similar Contents

灘本 明代† 田中 克己‡

Akiyo NADAMOTO Katsumi TANAKA

本論文では，2つの言語の異なる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 Bilingual Comparative Web Browser (B-CWB)を提案する．B-CWBによれば，ユーザは日本語と英語のニュースサイトを指定し，日本語のニュースサイトのページを順次閲覧するだけで，同時に類似する英語のニュースサイトのページを比較・閲覧することができる．B-CWBの特徴は，(1)ユーザのオペレーションに応じたコンテンツ同期自動提示（Content Synchronization），(2)Topic Structureを用いた類似ページの差異情報の発見である． B-CWBによりユーザは容易に自動で言語の異なる2つのWebサイトを比較することが可能となる．

We propose a new way of browsing bilingual web sites through concurrent browsing with automatic similar-content synchronization and difference-detection facilities. We call this system the Bilingual Comparative Web Browser (B-CWB) and it concurrently presents bilingual web pages in a way that enables their content of the Web pages to be automatically synchronized. The B-CWB allows users to browse two web news sites in a concurrently and compare the similar news articles written in different languages (English and Japanese). The major characteristics of the B-CWB are its content synchronization and difference detection: Content synchronization means that user operation (scrolling or clicking) on one web page does not necessarily invoke the same operations on the other web page to preserve similarity of content between the two web pages. Difference detection means that the B-CWB analyzes two similar web pages shown concurrently to discover the several "differences" between them. This facility is important in comparing two news articles that report the same affairs.

1. はじめに

現在，ラジオやテレビに続くメディアとしてWebは一般的となっている．Webサービスが現在のように一般的になる以前は，我々はニュース等の報道をテレビやラジオ，新聞などから取得していた．これらテレビや新聞で報道される情報は各々の国で編集されたものや，外国に発信することを目的とした各国の通信社から発信される情報がほとんどである．すなわち，諸外国で報道されているニュースそのものを，我々は自国にいながらテレビ番組や新聞から取得することが困難であった．しかしながら，現在，海外のニュースサイトのWebページを閲覧することによりその国のニュースをそのまま取得することが可能となった．ニュースソースは報道の視点や編集により種々の捉え方ができ，ひとつの事件や事故も国が異なれば視点が異なり，視聴者や読者は同じニュースでも異なった印象を持つ．例えば，A国とB国間の戦争に関するニュースでは，A国の報道ではB国の悪いところが誇張されるであろうし，またその逆も考えられる．しかしながら，ひとつのニュースソースが各国でどのように報道されているか公平に比較したい場合がある．この場合，各国のニュースサイトのWebページを閲覧することにより，その国で報道されているニュースを取得し，視聴者自身でどのように報道されているかを把握し比較することができる．このように，Webは我々の生活や世界観を大きく変えてきている．

現在，3600万以上のWebサイトがインターネット上に存在している[1]．これらWebサイトの中には1万ページ以上のWebページを持つサイトも多く，Webページは日々膨張している．複数のWebページを比較するには，これらのWebページの中から類似するページを探し出し，その類似したページを各々別のブラウザを開いて，各々のウィンドウをクリックしスクロールしコンテンツを読まなければならない．ましてや，言語の異なるサイトにおいては，複数のブラウザを同時にオペレーションするだけでなく，複数の言語を同時に理解しなければならず，ユーザがマルチリンガルでない場合，複数ページの閲覧はより困難である．そこで，我々は言語の異なる複数のWebページを同時に比較提示するブラウザがあると便利であると考え，Billingual Comparative Web Browser(B-CWB)を提案する．

これまで我々は同じ言語で記述されている複数のWebサイト内の類似Webページを同時比較するWebブラウザであるComparative Web Browser（CWB）[2]，[3]を提案してきた．CWBでは，ユーザが指定した複数のWebサイトから類似するWebページを発見するとともに，そのページの類似する部分も発見する．そして，その類似する部分を用いて，バック，フォワード，クリック，スクロールといったユーザのオペレーションに応じてコンテンツを同期させ自動提示する．本論文で提案するB-CWBの基本コンセプトは先に提案したCWBと同じであるが，以下の点が異なる．

- 言語の異なるWebサイトを比較提示する．
- CWBでは類似するページと類似するページの部分を発見したが，B-CWBでは類似するページ（または部分）だけでなく，類似するページの差異情報も発見する．

類似ページ及びそれらのページの差異情報の発見にtopic structureを用いる．B-CWBの特徴は以下のとおりである．

- ユーザの読んでいるWebページ（または部分）と類似するWebページ（または部分）の発見
- ユーザのオペレーション（バック，フォワード，クリック，スクロール）に応じたコンテンツ同期自動提示
- Topic Structureを用いた類似するWebページの差異情報の抽出

本論文ではページの部分をページの段落とする．

以下，2章ではB-CWBの基本コンセプトを，3章では類似ページの差異情報の発見を，4章でプロトタイプシステムについて述べ，5章でまとめについて述べる．

†正会員 独立行政法人通信総合研究所 nadamoto@crl.go.jp
‡正会員 京都大学大学院 情報学研究科社会情報学専攻
独立行政法人通信総合研究所
tanaka@dl.kuis.kyoto-u.ac.jp

2. 基本コンセプト

B-CWB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．

図１　画面イメージ

Fig.1　Picture of Display

基準サイトと比較サイト

B-CWBはユーザが指定した基準となるWebサイトとそれと比較したいWebサイトにより構成される．本論文では基準となるWebサイトを基準サイトと呼び，比較するWebサイトを比較サイトと呼ぶ．また，基準サイトを日本語のサイトとし，比較サイトを英語のサイトとする．そして，ユーザの指定した基準サイトのページを基準ページと呼び，比較サイトの類似するページを類似ページと呼ぶ．ユーザは基準サイトのページを閲覧するだけで，B-CWBは類似するページを比較サイトより検索し，同時に提示する．これにより，ユーザは言語の異なる2つのページを同時に閲覧，比較することが可能となる．図１では，ユーザは基準サイトに朝日新聞のサイト[4]を，比較サイトにCNNのサイト[5]を指定している．

視点差異情報の取得

B-CWBでは主題と内容からなるTopic Structureを用い，ユーザの指示により以下の種類のページの関連情報を取得する．内容相違・主題相違を指定することにより，ユーザは類似するページの差異情報を取得することが可能となる．

- 全体類似
  主題と内容両方が類似しているページまたは段落を示す．全体類似であるページ同士は同じニュースで同じ視点を持った記事である．
- 内容相違
  主題が類似しているが，内容が異なるページまたは段落を示す．内容相違のページは同じニュースであるが視点の異なる記事を示す．
- 主題相違
  内容が類似しているが主題が異なるページまたは段落を示す．主題相違のページは視点が同じであるが異なるニュースの記事である．

Content Synchronization

B-CWBではクリックする，スクロールする等のユーザのオペレーションに応じてコンテンツを同期させ提示するContent Synchronizationを行っている．以下にContent Synchronizationの機能を述べる．

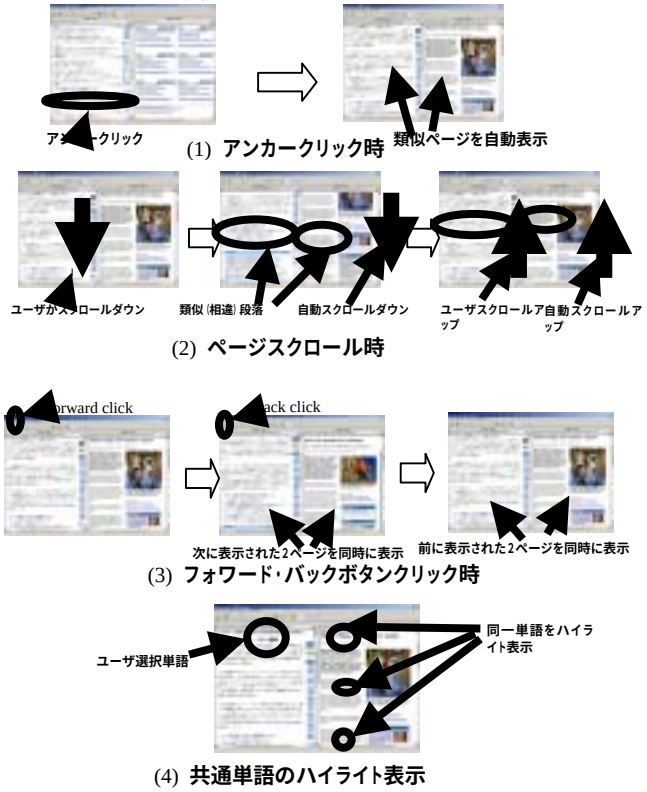

図２　インタフェイス例

Fig.2 Example of the Interface Picture of Display

- アンカークリック

図2(1)に示すように，ユーザは基準サイトとして日本語のニュースサイトを指定し，比較サイトとして英語のニュースサイトを指定すると，B-CWBは各々のサイトのトップページを表示する．次にユーザが基準サイトから見たいページのアンカーをクリックすると，B-CWBはそのページと類似するページを比較サイトから自動で発見し提示する．この時B-CWBはページレベルのContent Synchronizationを行っている．

- ページスクロール

ユーザが基準ページをスクロールした時，基準ページの表示ウィンドウの中央に表示された段落と関連する段落を，類似ページから抽出し比較サイトのウィンドウを自動でスクロールして提示する（図2(2)参照）．ここで言う関連する段落とは，上記視点差異情報の取得においてユーザが指定した関連情報を示す段落である．B-CWBは，ページスクロール時に段落レベルのContent Synchronizationを行っている．

- フォワード・バックボタン

ユーザがB-CWBのフォワードボタンまたはバックボタンをクリックしたとき，B-CWBは基準サイトのウィンドウと比較サイトのウィンドウ両方において，前のページまたは次のページを同時に提示する（図2(3)参照）．

- 単語の強調

B-CWBは2つのページを同時に提示しているため，ユーザが

特定の単語を発見することは困難である．そのため，ユーザが基準ページの単語を選択したとき，類似ページ上の同じ単語をハイライトして強調表示させる（図2(4)参照）．

差異情報の視覚化

B-CWBは類似するページの差異情報を取得した後，Content Synchronizationを行い，ユーザのオペレーションによって2つのWebページを同期させて提示する．この時，2つのWebページを同時に表示しているため，どの段落が類似段落なのか，または差異情報を持つ段落なのかユーザにとってわかりにくい．そこで，コンテンツの類似及び差異情報を視覚化する必要がある．視覚化する方法としてはいくつか考えられるが，本論文ではまず第一段階として，類似または差異情報を持つ段落を色分けをして提示する．これにより，ユーザに類似及び差異情報を持つ段落を明確に提示することが可能となる（図1参照）．実際には，全体類似を赤，内容類似を緑，主題相違を青として色分けを行う．

しかしながら，同じページに類似または差異情報を持つ段落が複数存在する場合がある．そのため，例えば主題相違の場合，最初の主題相違の段落は青，次の主題相違の段落は水色といったように，１ページに提示される関連情報を持つ段落ごとに色のグラディエーションをつけてわかりやすく提示する．

3. 類似ページの差異情報の発見

B-CWBはリアルタイムで2つのサイト間における類似ページを発見し提示しなければならない為，各ニュースサイトをあらかじめ1時間に1回クローリングしておき，ページ毎に単語の関連情報及び各ページの段落情報を格納するデータベースを作成する．このデータベースをキーワードデータベースと呼ぶ．以下の手順にて，ユーザの指定した関連情報に基づいた類似ページを比較サイトから検索する．

3.1 比較領域の抽出

B-CWBでは言語の異なるサイトを比較するため，単語の辞書ソフトを使用する．現在の単語の辞書ソフトは完全な言語変換を行えないため，我々はあらかじめ2つのサイトにおいて比較する領域を限定し類似ページを発見することを考えた．この比較する領域を比較領域と呼ぶ．ニュースサイトにおいては，分野ごとにカテゴリに分類されている．そこで，類似するカテゴリを発見することにより，比較領域を決定する．以下に比較領域の抽出方法を述べる．

- 各々のサイトをディレクトリページとコンテンツページに分ける．ここでいうディレクトリページとは，ページのほとんどがアンカーにより構成されているページを指す．ディレクトリページ以外のページをコンテンツページと呼ぶ．
- ディレクトリページを示すアンカーテキストをカテゴリ名とする．つまりは，基準ページのカテゴリ名は基準ページをリンクしているディレクトリページを指すアンカーテキストである．
- 基準ページのカテゴリ名を辞書ソフトを用いて英語に変換する．
- 比較サイトにおいてその基準ページのカテゴリ名を発見する．発見したカテゴリの子節点となるページすべてを含む領域が比較領域となる．

3.2 Topic Structureの生成

我々は，CWBでは類似ページを発見するために，ページのタイトル，サブタイトルから主題語を抽出しそれ以外の内容から内容語を抽出した．B-CWBでは，比較するサイトを作成する国が異なるため，タイトル，サブタイトル等のページ構成が類似しているとは限らない．そこで，松倉ら[6]が提案した単語の共起関係を用いたTopic Structureに基づいた方法を用いる．ページPにおけるTopic Structure TPはトピックt_i，i∈(1,...,n)からなり，t_i は主題語S_i と内容語の集合C_i のセットからなる．また，C_i は複数の内容語c_{im} ，m∈(1,...,k)で構成される．すなわちTPは以下のとおりである．

$$TP = \{t_1,...,t_i,...,t_n\}$$

$$t_1 = \{s_i, C_i\}$$

$$C_i = (c_{i1},...,c_{im})$$

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．B-CWB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

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

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．

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， ...， s_i，...， 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}，...， 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Topic Structureはt_i =(s_i, C_i),...， t_n=(s_n, C_n)となる．これらのTopic Structureはキーワードデータベースに格納される．

3.3 類似ページから差異情報の発見

B-CWBは上記で求めた主題語と内容語の特徴ベクトルからユークリッド距離を用いてページレベルと段落レベルの類似コンテンツの差異情報を取得する．この時，我々は辞書ソフトを用いて，英語の主題語と内容語を日本語に変換する．1つの英語には複数の日本語訳が対応している場合がある．その場合，複数の日本語を主題語の候補とする．2ページ（段落）間の主題語の類似度Sim_s，内容語の類似度Sim_cは以下のように決定される．

$$Sim_s = \sqrt{(f_b(s_1) - f_c(s_1))^2 + ... + (f_b(s_n) - f_c(s_n))^2}$$

$$Sim_c = \sqrt{(f_b(c_1) - f_c(c_1))^2 + ... + (f_b(c_m) - f_c(c_m))^2}$$

$$F_b = tf(s_n) \times weight(s_n)$$

ここで，$f_b(s_n)$は基準ページ（段落）の主題語の特徴ベクトルの要素であり，$f_c(s_n)$は比較するページ（段落）の主題語の特徴ベクトルの要素である．同様に，$f_b(c_m)$は基準ページ（段落）の内容語の特徴ベクトルの要素であり，$f_c(c_m)$は比較するページ（段落）の内容語の特徴ベクトルの要素である．上記のように求めた主題語の類似度，内容語の類似度に基づき2ページ（段落）間の関連情報を以下のように求める．

- 全体類似

主題語の類似度と内容語の類似度がある閾値以下の場合，その2つのページ（段落）は主題語と内容語が類似しているため全体類似のページ（段落）となる．

- 内容相違

主題語の類似度がある閾値以下であり，内容語の類似度は閾値以上である場合，その2つのページ（段落）は主題語が類似しているが内容語が相違しているとし，内容相違のページ（段落）となる．

- 主題相違

主題語の類似度がある閾値以上であり，内容語の類似度は閾値以下である場合，その2つのページ（段落）は主題語が相違しているが内容語が類似しているとし，主題相違のページ（段落）となる．

4. プロトタイプシステム

我々はMicrosoft C#を用いてB-CWBのプロトタイプシステムを作成した．キーワードデータベースにはOracle 9iを使用し，日本語の形態素解析にMeCab[7]を英語の品詞解析にBrill's Tagger[8]を使用した．また，日英，英日辞書とてEIJIRO[9]を用いた．図1にB-CWBの画面を，以下にB-CWBのシステムアーキテクチャを示す．

B-CWBはキーワードデータベースを含む類似検索を行うサーバと，ページを同期させて表示するクライアントからなる．

① 前処理として，サーバ側で日本語と英語のニュースサイトをあらかじめクローリングし，各々のサイトのすべてのページのTopic Structureと段落情報を抽出しキーワードデータベースに格納する．ここで格納される単語の言語はそのサイトの言語である．
② B-CWB起動後ユーザは基準サイトと比較サイトのURLを指定する．B-CWBは各々のサイトのトップページを表示する．我々のプロトタイプシステムでは，基準サイトを日本語のニュースサイトとし，比較サイトを英語のニュースサイトとした．
③ ユーザは基準サイトの閲覧したいページのアンカーをクリックする．
④ システムはキーワードデータベースからユーザの指定した基準ページのTopic structureを取得する．
⑤ システムは基準ページのカテゴリ名を抽出し，比較サイトの比較領域を決定する．
⑥ システムは基準ページと比較サイトの比較領域内にあるWebページ及びその段落間の類似度を求める．この時，辞書ソフトを用いて英語の名詞を日本語へ変換する．
⑦ 最も類似度が高いページを類似ページと決定し，基準ページと類似ページをウィンドウに表示する．
⑧ 基準ページの段落と類似する段落が類似ページになかった場合，システムは類似段落を他のページから検索する．そして類似段落が他のページにあった時，基準ページ内のその段落に部分類似アイコンを表示する．ユーザがこの部分類似アイコンをクリックした時，システムは類似段落を含むページを別のウィンドウで部分類似ページとして表示する．
⑨ ユーザが基準ページを操作した場合，システムはその操作に同期させて類似ページを表示する．
⑩ ユーザが次の基準ページをクリックした時，システムは④から⑨の操作を繰り返す．

5. まとめ

本論文では，言語の異なる2つのWebページを同時に比較提示するブラウザである，Billingual Comparative Web Browser(B-CWB)を提案した．B-CWBは，ユーザの指定した日本語のWebサイトから単語の出現頻度とその単語の共起関係からなるTopic Structureを抽出する．そのTopicStructureを用いて英語のWebサイトからユーザの指定した全体類似，内容語相違，主題語相違の3種類の関連情報に基づき，類似ページを発見し，同時に提示する．ユーザは日本語のWebサイトを操作するだけで，類似する英語のサイトのページが同期して自動で提示および操作されるため，これら2つのページを同時に閲覧することができる．

B-CWBの特徴は以下のとおりである．

- ユーザの読んでいるWebページ（または段落）と類似するWebページ（または段落）の発見
- ユーザの振る舞い（バック，フォワード，クリック，スクロール）に応じたコンテンツ同期自動提示
- Topic Structureを用いた類似するWebページの差異情報の抽出

ユーザはB-CWBを利用することにより，容易に自動で言語の異なるWebサイトの類似ページの差異情報を取得することが可能となる．

今後の課題として，以下の2点が考えられる．

- より明確な差異情報の視覚提示方法の提案
- 多言語複数サイトを対象としたB-CWBの提案

[文献]

[1] netcraft ホームページ
http://www.netcraft.com/survey/
[2] 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
[3] 灘本 明代, 田中 克己, “CWB:類似Webページの比較同期提示機能を有するブラウザの提案”, 日本データベース学会Letters, ISSN 1347-8915 VOl1, No2. pp.36-39 2003年3月
[4] 朝日新聞ホームページ http://www.asahi.com
[5] CNN ホームページ http://www.cnn.com
[6] 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.
[7] MeCab ホームページ
http://cl.aist-nara.ac.jp/~taku-ku/software/mecab
[8] Brill's tagger ホームページ
http://www.cs.jhu.edu/~brill/
[9] EIJIRO ホームページ http://www.alc.co.jp/

灘本 明代 Akiyo NADAMOTO
独立行政法人通信総合研究所勤務．2002 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）．マルチメディアコンテンツの情報配信，閲覧に関する研究に従事．情報処理学会，日本データベース学会会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授． 1976 年京都大学大学院前期博士課程修了，工学博士．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウエア科学会，情報処理学会，日本データベース学会会員．