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

CWB: A Web Browser with Functions of Comparison and Synchronized Presentation of Similar Web Pages

灘本 明代† 田中 克己‡
Akiyo NADAMOTO Katsumi TANAKA

本論文では，類似 Web サイトの Web ページを同時に比較表示するブラウザである Comparative Web Browser(CWB)を提案する．CWB によれば，ユーザは基準となる Web サイトと比較したい Web サイトの 2 つの Web サイトを指定する．次に，基準となる Web サイト内の閲覧したい Web ページを指定する．CWB はその指定された Web ページからキーワードとなる単語を特徴ベクトルを用いて抽出し，そのキーワードを用いて比較したい Web サイトから類似した Web ページを自動で発見する．そしてこれら 2 つの Web ページを同期させながら提示する．CWB の特徴はページ全体または部分の類似している Web ページを発見し，これらを同時に同期して提示することである． CWB を使用することにより，ユーザは基準となるサイトの Web ページを順次閲覧するだけで，比較したいサイトの類似ページを容易に閲覧することが可能となる．

In this paper, we propose a new Web browser which represents concurrently two similar Web pages in two similar Web sites. We call this browser a "Comparative Web Browser (CWB)". In order to use the CWB, first, users specify a basic Web site and a compared Web site. The CWB extracts a feature vector composed of keywords in the Web page at the basic Web site automatically. Next, the CWB finds the similar Web page using the keywords, and presents those Web pages concurrently. The notable functions of the CWB are automatic finding of a similar Web page, or a similar passage, and concurrent presentation of similar Web pages or passages according to user's behaviors. By using the CWB, users can browse a Web page together with a similar Web page of a different site in a comparative manner.

1. はじめに

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ページは膨大な数となっている．このように，Webサイトを持つことは企業や大学にとって当然のようになっている．それに伴い，一般社会[1]と同様，Webサイトはポータルサイトやニュースサイトのように類似した分野や業種で分類することが可能である．

しかしながら，この類似したWebサイトのページを比較しようとした場合，特定のWebページを比較しているサイト[2][3]で検索するか，Webサイトごとに閲覧したいページを各々検索して提示した後，人手による比較を行わなければならない．例えば，あるニュース記事をサイトごとにどのように記述されているか比較したい場合，各々のサイトを個々に別のブラウザで開き，その関連するページを各々提示して読まなければならない．これではユーザにとって，複数のサイトを比較することは困難である．そこで我々は，一つのWebサイトのページを提示している時，同時に他のサイトの類似したWebページを自動で提示するブラウザがあると便利であると考え，Comparative Web Browser(CWB)を提案する．

図1に示すように，CWBはユーザの指定した基準となるWebサイトとそれと比較するWebサイトにより構成される．ユーザは基準となるWebサイトと比較したいWebサイトの2つのWebサイトのURLを指定し，その基準となるWebサイト内の閲覧したいWebページを指定する．システムはそのWebページと類似したWebページを比較するWebサイトから自動で発見し，同時に提示する．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，ユーザが指定した基準サイト内のWebページを基準ページと呼び，基準ページと類似した比較サイトのWebページを類似ページと呼ぶ．

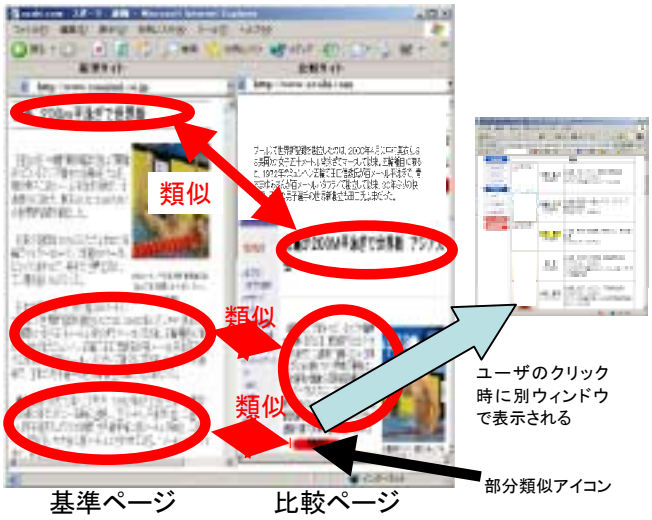

図1 CWB画面イメージ

Fig.1 Comparative Web Browsing

CWBでは，比較サイトから類似ページを自動で発見すると共に，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振る舞いに応じて基準ページと同期させて類似ページを提示する．ここでは，ユーザが操作するのは，あくまで基準サイト内のWebページに対してのみとする．また基準ページと類似ページのサイトが異なるため，基準ページの内容がすべて類似ページに含まれているとは限らない．つまりは，基準ページの内容が比較サイト内の複数のWebページにまたがっている場合がある．このような場合，内容が類似している部分を含むWebページも提示する必要がある．そこで，CWBではユーザが指定したWebページに最も類似しているWebページを同時に提示すると共に，その類似ページには含まれ

† 正会員 独立行政法人通信総合研究所 けいはんな情報通信融合センター 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 tanaka@dl.kuis.kyoto-u.ac.jp

ていない，基準ページと似た部分を持つWebページは別ウィンドウで提示することを行う．

CWBの特徴を以下に示す．

- 基準サイト内のWebページから問い合わせとなるキーワードを自動抽出する．
- 基準ページ全体と類似しているWebページを比較サイトから発見する．
- 基準ページの部分と類似しているWebページを比較サイトから発見する．
- ユーザの振る舞いに応じて，Webページ全体や部分が類似する箇所を抽出し基準ページと同期させ,類似ページを提示する．

CWBにより，ユーザは基準サイト内のWebページを順次閲覧するだけで，比較サイトの類似ページを容易に閲覧し比較することが可能となる．

以下，2章ではConcurrent Web Browserシステムの構成を，3章では類似ページ検索部を，4章ではインタフェース部について述べ，5章でまとめについて述べる．

2. Comparative Web Browser システム概要

CWBは類似ページ検索部とインタフェースの2つの機能からなる．類似ページ検索部は，基準サイト内のWebページからキーワードを自動抽出し，そのキーワードを用いて基準ページ全体または部分と類似しているWebページを比較サイトから発見する機能である．インタフェース部はユーザの振る舞いに応じて，Webページ全体や部分の類似する箇所を抽出しユーザに提示する機能である．

CWBの処理ステップを以下に示す．

① ユーザは基準サイトと比較サイトのURLを指定し，基準サイトから閲覧したいWebページを選択する．この時，基準サイトと比較サイトは類似した内容を持つサイトとする．
② ユーザが指定したWebサイトにおいてすべてのWebページの単語の出現頻度をあらかじめ求めデータベースに保存する．
③ 基準ページからキーワードを自動抽出する．
④ 抽出したキーワードを用いて比較サイトのWebページの類似度を算出する．
⑤ 類似度の最も高いページを類似ページとし，CWBウィンドウに提示する．
⑥ ユーザが現在表示している基準ページの段落と類似している類似ページの段落を発見し提示する．
⑦ 類似ページに類似している段落がない場合，基準ページの提示部分と類似している部分をもつ他のWebページを比較サイトから検索し，アイコンとして提示する．このアイコンを部分類似アイコンと呼ぶ．
⑧ ユーザが部分類似アイコンをクリックした時，そのWebページを別ウィンドウで提示する．
⑨ ユーザが次の基準ページを提示した場合，③から⑧を行う．
⑩ ユーザがウィンドウをバックまたはフォワードした場合，比較サイトに表示する類似ページもユーザの振る舞いに同期して変更する．
⑪ ユーザが基準ページの中からある単語を選択した場合，類似ページに含まれるその単語を強調して提示する．

上記ステップの①から⑥は類似ページ検索部であり，⑦から⑪はインタフェース部である．図2にシステム構成を示す．

図1の画面イメージはユーザは基準サイトと比較サイトに2つのニュースサイトを指定した場合の例である．ユーザが基準サイトのAニュースサイトの水泳の新記録に関するWebページを選択すると，システムはそのページからキーワードを抽出し，比較サイトのBニュースサイトから同じ水泳の新記録に関するWebページを発見する．そして，ウィンドウの左側にAニュースサイトのWebページを右側にBニュースサイトのWebページを提示する．このとき，Aニュースサイトの表示されているWebページでは選手の経歴が記述されているとし，Bニュースサイトの表示されているWebページに記述されていないとする．このとき，基準ページであるAニュースサイトの選手の経歴の記述の段落の下にアイコンが表示される．このアイコンをクリックすると図1の右部分に示すように，別ウィンドウでBニュースサイトの選手の経歴が記述された別のWebページを表示する．

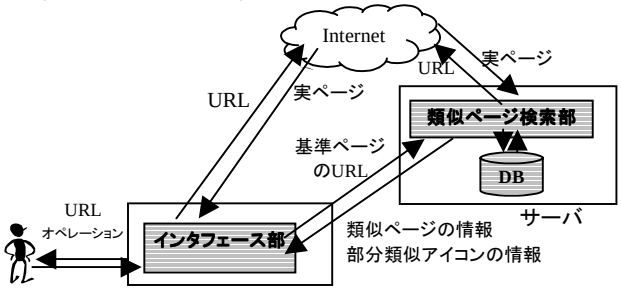

図 2　CWB システム構成図

Fig.2　System Architecture of the CWB

3. 類似ページ検索部

3.1 キーワードの抽出

Webページの特徴を捉え比較する研究[4][5][6]は多くあるが，tf/idfを用いた手法が多く用いられている．この手法は複数のWebページにおいては有効であるが，１つのWebページにおいては全体の単語数が少なく有効であるとはいいがたい．それに対し，小山ら[7]は，ユーザの入力したキーワードを，Webページの構造から主題となるキーワードと内容に含まれるキーワードに分類し検索精度を上げる研究を行っている．本論文では，小山らの手法に基づき，Webページの構造から主題となるキーワードと内容に含まれるキーワードを抽出する．ここでは，主題となるキーワードを主題キーワードと呼び，内容に含まれるキーワードを内容キーワードと呼ぶ．

Webページは，ページの見出しとなるタイトルがあり，それ以下いくつかの段落に分かれ，それぞれの段落はサブタイトルや本文で構成されるといった階層構造を持つ場合が多い．そこで我々は，このWebページの階層構造を用いてキーワードを抽出することを行う．実際には，タイトルやサブタイトルに含まれる単語はそのページ全体や部分を顕著に表す単語であると考え，主題を示すキーワードとする．また，そのタイトルやサブタイトルの内容を示す文章に含まれる単語はその内容を示すキーワードであると考え，主題キーワードと内容キーワードの2種類のキーワードの抽出を行う．以下にキーワードの抽出手順を示す．

① 構造タグを用いてWebページから木構造を作成する．
② Webページ内の単語の出現頻度を算出する．
　ここで対象にする単語とは名詞，固有名詞を示す．
③ 単語の出現頻度から各単語におけるベクトルを求める．
　名詞の品詞により重み付けを行い，各単語の出現頻度に

この品詞による重みを掛け各々の単語のベクトルを算出する．ここで，実験より重みは固有名詞3.0，数0.1，助数詞0.1，名詞一般1.0，その他名詞0.9とした．

④ タイトル，サブタイトルを発見する．
タイトル，サブタイトルはそれのみでタグで囲まれている単語または文である．また，他のWebページ内の文章と比較して文字サイズが大きかったり，強調されている場合が多い．そこで，単語または1文が<Font>タグまたは<H>タグで囲まれており，且つ文の最後が名詞，固有名詞で終了しているものをタイトル，サブタイトルの候補とする．タイトルはページの最初に出現する単語または文であり，木構造の最も浅く最も左に位置する単語または文である．サブタイトルはタイトル以外の候補となった単語または文である．タイトル，サブタイトルは入れ子構造になっている．

⑤ 主題キーワードを決定する．
タイトル，サブタイトルに含まれる単語を抽出し，その単語を各々のキーワードとする．この時，タイトル，サブタイトルは階層を持っているため，木構造を横型探査してキーワードを決定する．また，一つのタイトル，サブタイトルの中に含まれる名詞，固有名詞の単語をすべてキーワードとすると，主題キーワードが膨大な数になる場合がある．そのため，キーワードとなる単語は，単語のベクトルがある閾値α以上のものとする．タイトルのキーワードT_i，サブタイトルのキーワードST_{xk}は主題キーワード$inTitle$とする．ここでiはタイトルのキーワードの個数を示し，xはサブタイトルの個数を，kは一つのサブタイトルにおけるキーワードの個数を示す．つまりは$inTitle$は，

$$inTitle = (T_i, ST_{1j}, \ldots, ST_{xk})$$

となる．

⑥ 内容キーワードを決定する．
タイトル，サブタイトル以外の文章は内容を示すと考え，その内容キーワードを抽出する．基準ページの部分ごとの類似を求めることを考え，内容キーワード$inText_i, i \in \{1, \ldots, n\}$は基準ページの段落ごとに決定する．また，$inText_i$ は単語のベクトルがある閾値α以上のものとする．ここでいう閾値αは主題キーワードの単語ベクトルの閾値αと等価とする．ここでiは段落の番号を示す．つまり内容を示す文章に含まれ且つ単語ベクトルがα以上の単語を$C_i, i=1,2,\ldots n$とすると$inText_i$ は

$$inText_i = (C_0, C_1, \ldots, C_n)$$

となる．ここで$inText_i$ のキーワードの順位は単語ベクトルの大きい順とする．

3.2 類似ページの検索

抽出されたキーワードを用いて，比較サイトから類似ページを検索する．CWBでは基準ページと全体が類似しているWebページ，部分が類似しているWebページを取り扱う．ここでいうWebページの部分はWebページの段落を示す．Webページの段落は構造タグを用いたWebページの木構造の一つの節点となる．よって本論文ではWebページの木構造の一つの節点を単位として類似検索を行う．Webページの全体が類似しているページとは，類似した節点を最も多く持つページとする．

前節で求めた主題キーワードと内容キーワードを用いて比較サイトから基準ページに類似している類似ページを決定する．小山らの研究により，主題と内容に含まれるキーワードの意味が異なることが実験により実証されている．よって，本論文では，主題キーワードは比較サイトのWebページのタイトル，サブタイトルから検索を行い，内容キーワードは比較サイトのWebページ内の内容を示す文章から検索をする．しかしながら，サブタイトルがWebページ内に存在する場合と存在しない場合とでは，Webページの構造が異なる．そこで，各々の場合に分けて検索を行う．

a) Webページにサブタイトルがある場合

この場合，構造化されているWebページといえる．図3左図に示すように，タイトル，サブタイトルの子節点がその内容を示す文章となる．そこで，Webページにサブタイトルがある場合，サブタイトルとその子節点をひとつの固まりと考え，以下の手順で検索を行う．

1. 比較サイトのWebページのタイトル，サブタイトルから主題キーワードと類似しているものを検索する．
   タイトル，サブタイトルは入れ子構造になっているため，木構造の横型探査を行う．タイトル，サブタイトルが主題キーワードと類似していれば，その子節点である内容も類似していると考える．よって，主題キーワードと類似しているタイトル，サブタイトルの子節点の検索は行わない．ここではユークリッド距離を用いて類似度を求める．つまりは，主題キーワードの特徴ベクトルと最も距離が小さいタイトル，サブタイトルとその子節点を類似段落とする．
2. 内容文章から内容キーワードと類似しているものを検索する．
   タイトル，サブタイトルに主題キーワードが含まれている節点の子節点以外の節点の文章から内容キーワードと類似しているものを検索する．つまりは，内容キーワードの特徴ベクトルと最も距離が小さい節点を類似段落とする．

b) Webページにサブタイトルがない場合

この場合，Webページはサブタイトルによって構造化されていないWebページといえる．図3右図に示すように，タイトルをルートとし，それ以外の節点は内容を示す文章である．この場合，すべての節点を横型探査を行い，内容キーワードと類似している節点を検索する．つまりは，内容キーワードの特徴ベクトルと最も距離が小さい節点を類似段落とする．

このように，比較サイトのWebページごとに基準ページとの類似段落を発見する．ここで，類似段落が最も多いWebページが類似ページの候補となる．類似段落が最も多いWebページが複数ある場合，類似サイトにおけるリンク木の階層が最も浅く且つ最も左側に位置するページを類似ページとする．

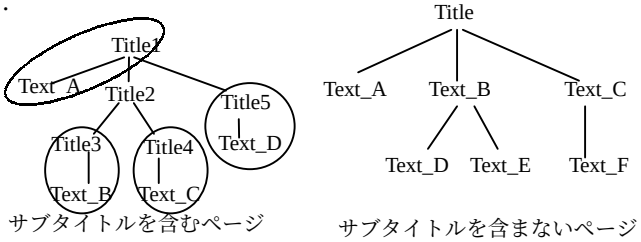

図3 Webページ内のタイトルの構造

Fig.3 Structure of Titles in a Web page

3.3 基準ページと類似ページの差分情報の取得

基準ページに含まれる内容がすべて類似ページに含まれ

るとは限らない．基準ページに含まれ且つ類似ページに含まれない情報が比較サイトの他のページにある場合がある．そこで，CWBではこの基準ページと類似ページとの差分情報を持つWebページを別ウィンドウで提示することを行う．

前節では比較サイトのWebページにおいて，段落ごとに基準ページとの類似段落を検索した．そこで，類似ページに類似段落を持たない基準ページの段落に含まれるサブタイトルのキーワードSTx_jまたは内容キーワード$inText_i$の類似度が最も高い段落を類似ページ以外の比較サイト内のWebページより発見する．その段落を持つWebページが基準ページと類似ページの差分情報を持つWebページとなる．差分情報を持つWebページの候補が複数ある場合，比較サイト内のリンク木の階層が最も浅く且つ最も左側に位置するWebページを差分情報を持つWebページとする．

4. インタフェース部

インタフェース部では，ユーザがウィンドウをクリック，スクロール，バック，フォワードなどのオペレーションを行うと同時に，基準サイトのウィンドウに提示されるWebページに同期させ，類似サイトのウィンドウに類似したWebページを提示する機能である．インタフェース部ではユーザの振る舞いに応じて以下の機能を提供する．

a) クリック時の類似ページの提示

ユーザが基準ページのアンカーをクリックしたとき，リンク先ページを新たな基準ページとする．そして，新たな基準ページからキーワードを抽出し，比較サイトから類似ページを発見して，基準ページと同期させて提示する．この時，基準ページと類似ページの差分情報を持つWebページを類似ページにアイコン化して提示する．

b) スクロール時の類似ページの類似部分の提示

1ページが長いWebページが多く存在する．この場合，ユーザはウィンドウをスクロールしてこのWebページを閲覧する．そこで，CWBではユーザが基準ページをスクロールして現在ウィンドウ上に表示されている基準ページの段落と類似している類似ページ内の段落を自動スクロールしてユーザに提示する（図4参照）．この時，類似ページに類似した段落がない場合，ユーザは類似部分アイコンをクリックすることにより，別ウィンドウで基準ページと類似ページの差分情報を持つWebページを閲覧することができる．

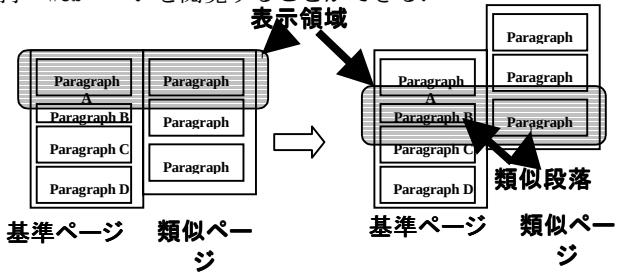

図 4 スクロール時の類似段落提示

Fig.4 Scrolling a Page with Passage-level Content Synchronization

c) バック，フォワード時の類似ページの提示

ユーザがブラウザのバックまたはフォワード機能により前または後ろのページを再度閲覧するとき，基準ページと類似ページを同期させて提示する．

d) 基準ページの単語の選択時の類似ページの提示

CWBではユーザは２つのWebページを同時に閲覧しているが，その場合，一目で類似情報を取得しにくいと考えられる．そこで，ユーザが基準ページ内で選択した単語を比較ページでも提示することにより，直感的に類似情報を取得できるようにする．

5. まとめ

本論文では，類似したWebサイトのWebページを同時に比較表示するブラウザであるComparative Web Browser(CWB)を提案した．CWBでは，ユーザにより指定された基準サイト内の基準ページからキーワードを抽出し，そのキーワードを用いて類似ページを比較サイトから自動で発見し，同時に提示する．発見するキーワードは主題キーワードと内容キーワードからなり，これらキーワードは各々タイトルサブタイトル検索および内容検索に用いた．そして，Webページの構造からなる木構造を用いて，類似ページを発見した．

CWBを使用することにより，ユーザは基準となるサイトのWebページを順次閲覧するだけで，比較するサイトの類似ページを容易に閲覧することが可能となった．

[謝辞]

本研究の一部は，平成14年度受託研究（独立行政法人通信総合研究所）「マルチメディアコンテンツのクロスメディア連携に関する研究」（代表：田中克己）による．ここに記して謝意を表します．

[文献]

[1] netcraft ホームページ
http://www.netcraft.com/survey/
[2] Gomez ホームページ http://www.gomez.com
[3] kakaku.com ホームページ http://www.kakaku.com
[4] B. Liu, K. Zhao, and L. Yi, "Visualizing Web Site Comparisons", The 11th International World Wide Web Conference (WWW2002), Honolulu, Hawaii, May 2002 (http://www2002.org/CDROM/refereed/571/index.html)
[5] M. Perkowitz, O Etzioni, "Towards adaptive Web Sites: Conceptual framework and case study", Artificial Intelligence 118, pp.245-275, 2000.
[6] I. Cadez, D. Heckerman, C. Meek, P. Smyth and S. White, "Visualization of Navigation Patterns on a Web Site Using Model Based Clustering", Technical Report, University of California, March 2000.
[7] 小山 聡，田中 克己，“話題の階層構造を反映したWeb検索手法の提案”，情報処理学会研究報告，Vol.2002, No.67 2002-DBS-128, pp.465-472 2002年7月

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．