デジタル・ドキュメント 28－2
（２００１．５．２５）

Webコンテンツ間の共通レイアウト自動解析

福田 健太郎　高木 啓伸　前田 潤治　浅川 智恵子

日本アイ・ビー・エム株式会社 東京基礎研究所
〒242–8502 神奈川県大和市下鶴間 1623–14
Tel: 046–215–4659　Fax: 046–274–4282
E-mail: kentarou@jp.ibm.com

あらまし

近年，Webを閲覧するための手段として，PDA・携帯電話などの小画面デバイスや音声ブラウザなどが一般に用いられつつある．通常のWebページはデスクトップコンピュータの画面サイズを考慮し，また視覚的に操作性を向上させるようなレイアウトになっているため，小画面デバイスや音声を用いた閲覧のためにはレイアウトの変更などが必要となる．この様なコンテンツの変換を行うためには，ページの構造や各部位の重要度などを記述したアノテーションを付加することが有効である．しかし，ニュースサイトなどは膨大な数のページを持つため，それぞれに詳細なアノテーションを付加する為には大変な労力が必要になる．一方で，これらのサイトでは同一のレイアウトを用いているページが多数存在するという特徴もある．そこで，本稿ではHTML文書のタグ構造およびその特徴値に基づいてページ間の距離を導出し，同一レイアウトを用いているページ群を自動的に検出する手法の提案を行う．本手法を用いる事により，同一レイアウトのページ間におけるアノテーション共有が可能となり，効率よいアノテーション付加が実現される．

Common layout extraction from Web pages

Kentarou FUKUDA　Hironobu TAKAGI　Junji MAEDA　Chieko ASAKAWA

IBM Japan, Ltd., Tokyo Research Laboratory
1623–14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan
Tel: +81 46 215 4659　Fax: +81 46 274 4282
E-mail: kentarou@jp.ibm.com

Abstract

In these days, people access to the Web by using various devices and methods, such as PDA, cellular phone and voice-based browsers. However, most Web contents are designed for desktop computers. Therefore, already-existing Web contents should be transcode to be suitable for each access devices and methods. For this purpose, some annotation-based transcoding systems have been developed. One of the most difficult problem of annotation is the cost to annotate Web contents. Many of popular sites, such as news sites, have a great number of Web pages and make new contents continually. Hence, it is almost impossible to annotate every contents in these sites. To solve this problem, we introduce the method to extract common layout from web pages. We focus on the structure and characteristics of particular HTML tags, which affect the layout of Web pages. Our method calculates the distance of web pages based on it. In the case where the distance is below the threshold, these pages can be considered as same layout pages. By using this method, a certain annotation can be applied to any Web pages that have identical layout. Therefore, the cost of annotation will be reduced.

1 はじめに

近年，World Wide Web(以下 Web) の役割の多様化により，情報の発信・収集のみならず，電子商取引，遠隔教育，コミュニティの形成など，様々な分野においてWebの重要性が増してきている．同時に，Webを閲覧するための手段も多様化してきている．

従来，Webにアクセスする場合にはパーソナルコンピュータ等を用いて，比較的大きな表示装置に内容を表示し，操作を行っていた．ところが，近年，携帯性・即時性に優れたPDA(Personal Digital Assistants) や携帯電話端末などを用いたWebへのアクセスが急速に広がりつつある．また，印刷文書へのアクセスが困難な視覚障害者にとって，合成音声を用いた音声ブラウザ等の読み上げソフトを用いたWebへのアクセスは重要な情報源となっている [1, 2]．さらに，音声による情報へのアクセスは，コンピュータに慣れ親しまない人々とのヒューマンインターフェースの向上や，表示装置が小さい，もしくは表示装置を持たないような小型コンピュータの補助的な出力手段としても期待される．

このようなWebへのアクセス手法の多様化に伴って，コンテンツをユーザの環境や要求に合わせて変換するトランスコーディングの研究が活発に行われている [3-10] ．

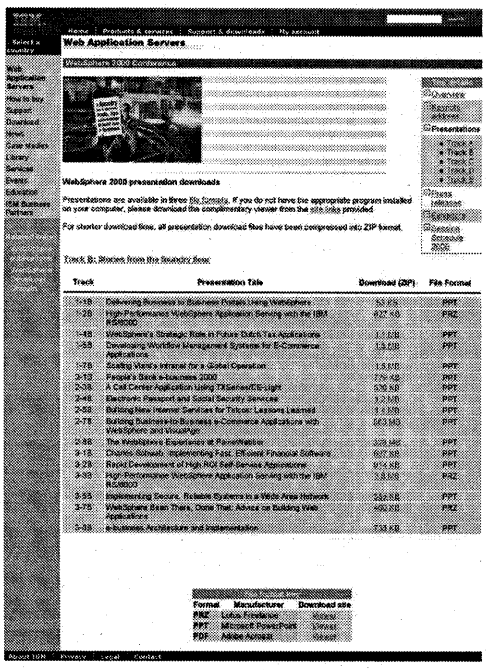
例えば，PDAや携帯電話などの機器では，画面サイズに制約があるため，デスクトップコンピュータの画面サイズを前提にデザインされたWebページを効果的に表示するためには，機器に応じたレイアウトの変更や画像の縮小などの変換が必要となる [3]．

また，Webページの多くは，目的の異なる情報を背景色やレイアウトテーブルなどの視覚表現を用いてグループ化・配置することにより，情報の集積度や視覚的な操作性の向上を図っている [5, 9, 11]．しかし，タグ順序に従って読み上げを行う音声ブラウザを用いた場合，これらの視覚的な情報へアクセスすることは非常に困難である [5, 11]．さらに，音声ブラウザによりWebにアクセスする場合，ページの先頭に配置された広告やリンクリストなどを順に読み進めるため，必要な情報へなかなかたどり着けないという問題も発生している [5-7]．

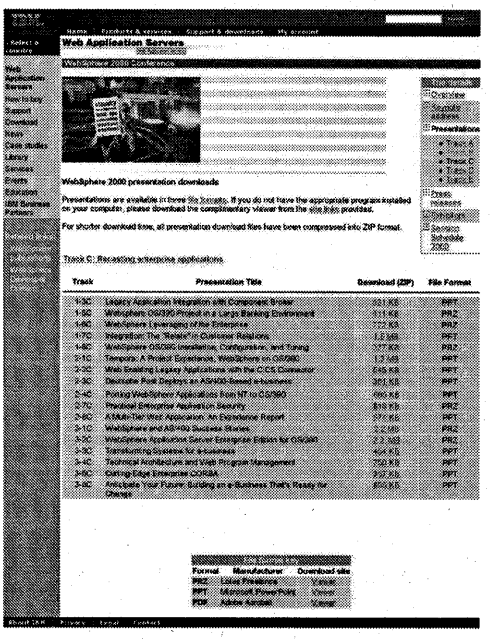
音声ブラウザを用いて必要な情報へ簡単にアクセスする為の方法として，過去のページや近隣ページとの差分を取ることにより，コンテンツの更新された部分のみを抽出する手法が研究されている [6, 7]．これらの方法により，Webページから，広告やリンクリスト等のテンプレートが取り除かれ，ニュースの記事や検索結果といったユーザが最も必要としている情報のみを読み上げることが可能となる．しかしながら，当該ページと過去・近隣ページのレイアウトが異なっている場合には，更新情報の抽出が出来ないことや，アクセス毎に動的に挿入される広告などは取り除けないこと，近隣ページへのリンクや図表のタイトルなど更新されていないが重要な情報なども取り除かれてしまう等の問題点があり，さらに精度の高いコンテンツの変換手法が求められている．

より高度なコンテンツの変換を実現するための手法として，アノテーション [3, 10, 12] を用いたトランスコーディングの研究が活発に行われている [3, 5, 9, 10] ．アノテーションとは，Web上のコンテンツのそれぞれに付与された言語的な付加情報である．例えば，文献 [5] においては，コンテンツを意味的なグループに分割し，その役割などをXML形式 [13] で保存している．このアノテーションに基づいてトランスコーディングを行うことにより，意味的な塊を崩したり，情報を欠落させたりすることなくWebページのレイアウトを変換することが可能になっている．さらに，各グループを重要度順に並べ替えることにより，音声ブラウザを用いた場合でも，広告やリンクリストに煩わされることなく，簡単に重要な情報へアクセスすることが可能となる．

このように，各コンテンツにアノテーションを付加することにより，高精度なトランスコーディングが可能となるが，ニュースサイトやデータベースサイトなどは膨大な数のページを持つため、それぞれに詳細なアノテーションを付加する為には大変な労力が必要になるとい



www.ibm.com/software/webservers/ws2000_present_b.html



www.ibm.com/software/webservers/ws2000_present_c.html

図 1: Web ページのレイアウトの例

う問題がある。また，ニュース記事などの場合は，頻繁に新規コンテンツ作成されている場合が多く，その更新にあわせてアノテーションを付加することはほぼ不可能である．

一方で，これらのサイトにおいては，ほとんどのページがテンプレートを用いて作成されており，同一のレイアウトを持つページが大きな部分を占めている．従って，同一のレイアウトを持つコンテンツ間でアノテーションの共有が可能となれば，アノテーション付加の効率を大幅に向上できると考えられる．さらに，新規に作成されたコンテンツのレイアウトが，既存のレイアウトと同一であった場合，新たにアノテーション付加を行う必要が無くなる．従って，更新頻度の高いニュースサイト等へのアノテーション付加も容易なものとなる．

そこで，本稿では Web ページのタグ構造を基に，同一のレイアウトを用いているページ群を自動的に検出するための手法を提案する．提案手法では，HTML 文書のタグ構造を解析し，レイアウトを決定する要因となるタグおよびその特徴値を列挙する．次にそれらの情報に基づいて，ページ間の距離を導出し，一定の条件を満たしたものを同一レイアウトを用いたページと判定している．本稿では，既存の Web サイトに提案手法を適用し，その有効性を示す．さらに，提案手法を用いて効率良くアノテーション付加を行うための手法についても検討を行う．

2 共通レイアウト解析手法

多くの Web サイトにおいては，目的の異なる情報を背景色やレイアウトテーブルなどの視覚表現を用いてグループ化・配置することにより，情報の集積度や視覚的な操作性の向上を図っている [5, 9, 11]．一般に，記事や検索結果などメインとなるコンテンツをページの中心に配置し，上方には広告やタイトル・日付などのヘッダー情報，ページの左右にはサイト内外へのリンクリストや関連情報，ページの下部には著作権情報などのフッター情報が配置されていることが多い (図 1)．これらのページの多くは，更

新された記事や検索結果をテンプレートに挿入するなどして自動作成されており，多くのページで共通のレイアウトが用いられている．

このような共通のレイアウトが用いられたページにおいては，コンテンツの意味的なグループやその役割なども共通であるため，レイアウト間でアノテーションを共有することにより，アノテーション付加の効率を向上することが可能であると考えられる．

本稿では，複数の Web ページから共通のレイアウトを用いているページを検出するための手法として，HTML 文書のタグ構造およびその特徴値を基にページ間の距離を導出する方式の提案を行う．

2.1 レイアウトタグと特徴値

Web ページにおけるレイアウトを決定する要因として，水平線やテーブルなどを記述する以下のようなタグの構造が挙げられる．

- 水平線 (HR)
- テーブルエレメント (TABLE, THEAD, TBODY, TFOOT, TR, TH, TD)
- スタイルコンテナ (DIV)
- 段落 (P)

以降では，これらのタグをレイアウトタグと呼ぶ．多くのサイトにおいて視覚的なレイアウト構築は，テーブルの入れ子により実現されており [14]，レイアウトタグの構造を明らかにすることで，共通レイアウトを導出することが可能になると考えられる．

そこで，提案方式においては，まず HTML 文書を解析しレイアウトタグの列挙を行う．この際，各レイアウトタグは HTML 文書内での位置が一意に明らかとなるように XPath [15] を用いて記述する．図 2 は，HTML 文書と列挙されたレイアウトタグのリストの例である．図中の矢印は HTML 文書内のタグと XPath との対応を表している．

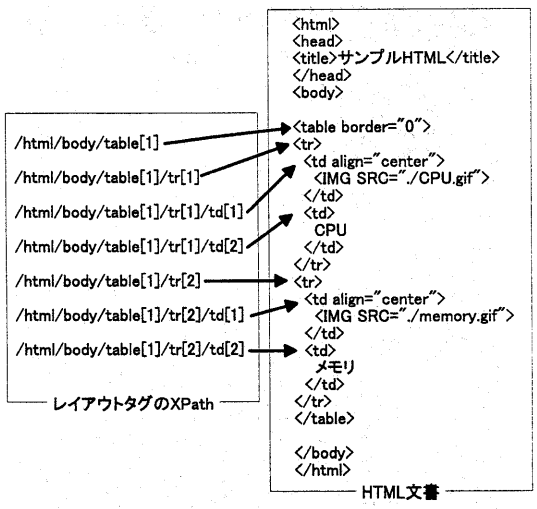
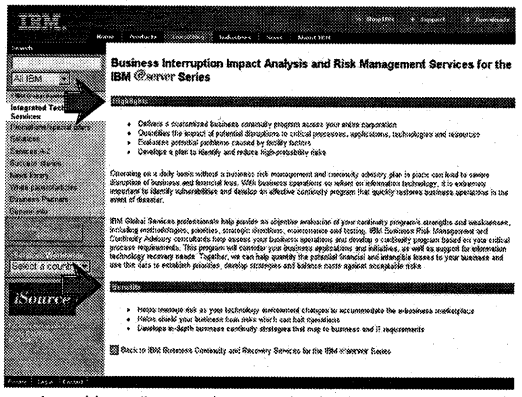


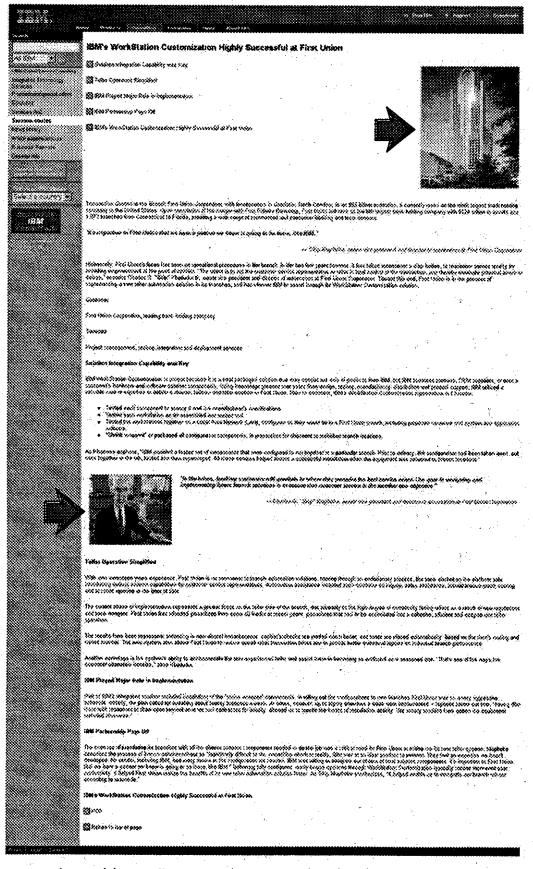
図 2: HTML 文書と列挙されたレイアウトタグ

図 2 の様に列挙されたレイアウトタグの一覧は，Web ページの概要を表すことになるが，この一覧が等しくても Web ページのレイアウトが等しいとは限らない．例えば，図 3 の 2 つのページはレイアウトタグのリストは全く同じであるが，レイアウトは大きく異なっている．図中，矢印で示した部分はそれぞれ対応するレイアウトタグ (TD) の位置だが、左のページの TD においては幅 (width) や背景色 (bgcolor) などを設定した上で文字が配置されているのに対し、右のページでは幅や背景色の指定無しに画像が配置されているのみである。このように，Web ページのレイアウトは，レイアウトタグの位置に加えて，レイアウトタグの利用方法によって大きく変動することがわかる．

そこで，本稿ではレイアウトタグの列挙の際に，それぞれのタグが持つ属性 (attribute) および，当該タグのサブツリー内に含まれるエレメント等の情報をその特徴値として関連付ける．例えば，レイアウトタグ TD における属性としては align, valign, bgcolor, colspan, rowspan, height, width が有り，さらに TD セル内のテキストのサイズや画像の数などが特徴値として挙げられる．このように，レイアウトタグに加えて特徴値を用いることで，Web ページのレ



http://www.ibm.com/services/its/us/bcintalert.html



http://www.ibm.com/services/its/us/firstunion.html

図 3: レイアウトタグのリストは等しいが，レイアウトは異なるページの例

イアウトを詳細に記述することが可能となる．

次節では，レイアウトタグおよび特徴値を用いて Web ページ間のレイアウトの相違を数値化するための手法について述べる．

2.2 ページ間距離の導出

前節で定義したレイアウトタグの位置情報および特徴値を比較することにより，同一のレイアウトを利用しているか否かの判定を行うことが可能になる．本稿では，Web ページ間のレイアウトの相違をページ間距離として数値化し，その距離に基づいて共通レイアウトを用いているか否かの判定を行う．

提案手法においては，距離を計算する 2 つのページに含まれるレイアウトタグの集合 A, B に対してページ間距離 D を以下のように定める。

$$D = \sum_i d_i(T_i) \tag{1}$$

式 (1) 中，T_i は $A \cup B$ を満たすレイアウトタグの集合の i 番目の要素である．また，d_i はレイアウトタグ T_i に関する距離関数であり，以下の式で与えられる．

$$\begin{aligned} d_i(T_i) &= W_i * \textstyle\sum_j \{W_{C_{ij}} * f_{C_{ij}}(C_{A_{ij}}, C_{B_{ij}})\} \\ &\qquad if(T_i \in (A \cap B)) \\ &= W_i * L_i \qquad otherwise \end{aligned} \tag{2}$$

ここで，W_i はレイアウトタグ T_i の重み係数，$W_{C_{ij}}$ は T_i における j 番目の特徴値 C_{ij} の重み係数である．$f_{C_{ij}}$ は，A, B における特徴値 C_{ij} の値 $C_{A_{ij}}, C_{B_{ij}}$ から特徴値間の距離を導出する関数で，特徴値が数値の場合には $|C_{A_{ij}} - C_{B_{ij}}|$ の単調増加関数として与える．特徴値が数値以外の場合には，特徴値が等しければ 0，異なる場合は正の値を返す二値関数とする．また，L_i はレイアウトタグ T_i が A, B のどちらか一方のページにしか存在しない場合の距離定数である．これらのパラメータは，対象となる Web サイトの特徴に合わせて設定する必要がある．例えば，図 1,3 のように背景色により意味的なグループを区分しているサイトの場合，特徴値

表 1: レイアウトグループ検出結果

	レイアウトグループ数	レイアウトグループに属さないページ数
LYCOS ニュース	216	98
朝日新聞社	884	2384

bgcolor の重みを大きくすることにより，共通のレイアウトを用いているページを効率よく検出することが期待できる．

提案手法においては，式 (1) を用いて導出されたページ間距離が一定の閾値を下回った場合，ページのレイアウトが同一であると判定する．また，ページ間距離に基づいてクラスタリングを行うことにより，多数のページから同一レイアウトを用いているページ群 (以下，レイアウトグループ) を一度に検出する事も可能となる．

このようにして得られたレイアウトグループに対し，レイアウト ID を割り当て，アノテーションをレイアウト ID に関連付けることで，共通レイアウトを用いている Web ページ間でアノテーションを共有することが可能となる．

3 評価実験

本章では，実際の Web コンテンツに提案手法を適用し，その有効性を明らかにする．実験の対象として LYCOS ニュース (http://news.lycos.co.jp) および朝日新聞社のサイト (http://www.asahi.com) から任意のページを取得し (それぞれ 25672, 19285 ページ)，それらのページからレイアウトグループの検出を行った．

実験に際しては，レイアウトタグとして HR, TABLE, THEAD, TBODY, TFOOT, TR, TH, TD の 8 種類のタグを用い，特徴値としてはそれぞれの属性を用いた．式 (2) における重み係数は全て 1 とし，レイアウトタグが一方のページにしか存在しない場合の距離定数は TABLE タグに関しては 10，それ以外のレイアウトタグに関しては 5 とした．また，簡単のため全ての距離関数を，特徴値が一致するときは

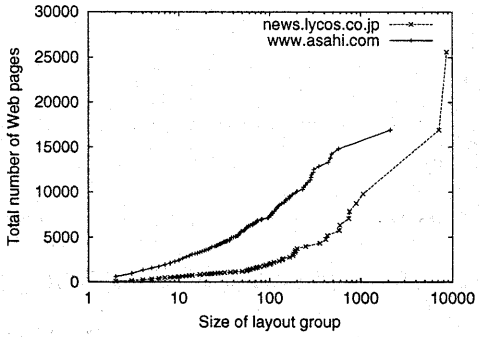


図 4: レイアウトグループのサイズと累積ページ数

0，異なるときは 1 を返す二値関数とし，ページのレイアウトが同一であるか否かの閾値としては 10 を用いる．

それぞれのサイトにおけるレイアウトグループ検出の結果を表 1 に示す．また，図 4 はレイアウトグループのサイズと累積ページ数の関係を表したものである．

表 1 の結果から，本稿で提案する手法を用いることにより，アノテーションのコストを大幅に削減できることが明らかである．例えば，LYCOS ニュースにおいては，約 300 種類のアノテーションを作成することで，サイト全体にアノテーション付加することが可能になる．ここで，同一レイアウトのページが存在しなかったページが 98 ページ残っているが，その殆どが特報などのページである．これらのページでは任意の位置に画像やタイトルを挿入する目的でテーブルが追加されたりするため，テンプレートとは異なるレイアウトになっている．

一方，朝日新聞社のサイトでは，レイアウトグループに属さないページ数や細かなレイアウトグループが多くなっている (表 1，図 4)．こ

のような結果になる原因としては，朝日新聞社のページにおいては，記事の中で比較的複雑なテーブルを用いていることが多いため，テンプレート部分が共通であっても，別レイアウトと判定されてしまうことが挙げられる．この様なサイトに対しては，テーブルの入れ子の深さに応じて，レイアウトタグの重みを小さくしたり，同一レイアウトか否かを判定する閾値を大きな値に設定することにより，効果的なレイアウト分割が可能となる．

また，オリンピックの種目別の結果をまとめたページ (www.asahi.com/olympic2000/wnpa以下のページ) などのように，入れ子の最深部に位置するテーブルは，視覚的なレイアウトを決定するためではなく，本来の意味でのテーブルとして用いられている場合もある．同様に，野球などのスポーツのスコアや画像の一覧などの場合も，テーブル全体が一つの意味的なグループとなっており，その構造の相違はあまり意味をなさない．このようなページに対しては，該当するテーブル内のレイアウトタグに対する重み係数を小さくすることにより，意味的に同一のレイアウトであるページを，異なるレイアウトであると判定されることを防ぐことができる．テーブルが本来の意味でのテーブルであるか否かを判定するための指標としては，テーブルのセル内のエレメントに関する特徴値を用いる方法が有効であると考えられる．セル内に他のテーブルを持つ場合にはレイアウトを決めるためのテーブルである場合が多く，比較的短いテキストや単一の画像のみが含まれる場合には本来のテーブルである確率が高い．

4 まとめ

本稿ではHTML文書のタグ構造およびその特徴値に基づいてページ間の距離を導出し，同一レイアウトを用いているページ群を自動的に検出する手法の提案を行い，実際のWebサイトに適用することでその有効性を示した．提案方式を用いることにより，同一レイアウトのページ間でアノテーション共有が可能となり，効率よいアノテーション付加が実現される．

今後の課題としては，Webサイトやページの特徴に適したパラメータの決定方法に関する検討を行う必要がある．また，提案手法を用いて検出したレイアウトグループの分割・統合などの操作をユーザが行えるよう拡張するとともに，ユーザの操作に応じて自動的にパラメータを修正する手法に関する検討も行う．さらに，テンプレート部分など，ページの一部のレイアウトが共通であるページ群を検出し，その共通部分に関するアノテーションを共有することにより，アノテーションの効率をさらに高めることが出来ると考えられる．

参考文献

[1] C. Asakawa and T. Itoh, "User interface of a home page reader," *Proceedings of ACM ASSETS '98*, pp. 149–156, April 1998.

[2] 渡辺 哲也, 坂尻 正次, 指田 忠司, 岡田 伸一, "視覚障害者の windows パソコン利用状況調査," **電子情報通信学会技術研究報告** (SP2000-47), pp. 37–42, August 2000.

[3] M. Hori, G. Kondoh, K. Ono, S. ichi Hirose, and S. Singhal, "Annotation-based web content transcoding," *Proceedings of 9th International World Wide Web Conference*, pp. 197–211, May 2000.

[4] H. Takagi and C. Asakawa, "Transcoding proxy for nonvisual web access," *Proceedings of ACM ASSETS 2000*, pp. 164–171, November 2000.

[5] C. Asakawa and H. Takagi, "Annotation-based transcoding for nonvisual web access," *Proceedings of ACM ASSETS 2000*, pp. 172–179, November 2000.

[6] 高木 啓伸, 浅川 智恵子, "視覚障害者のためのトランスコーディングシステムにおける WEB ページシンプリフィケーショ

ン手法," 電子情報通信学会 技術研究報告 (WIT00-23), pp. 67–72, August 2000.

[7] T. Ebina, S. Igi, and T. Miyake, "Fast web by using updated content extraction and a bookmark facility," *Proceedings of ACM ASSETS 2000*, pp. 64–71, November 2000.

[8] 前田 潤治, 小林 真, "ユーザの視覚特性に適応したウェブページの変換," 電子情報通信学会 技術研究報告 (WIT00-22), pp. 61–66, August 2000.

[9] A. W. Huang and N. Sundaresan, "A semantic transcoding system to adapt web services for users with disabilities," *Proceedings of 9th International World Wide Web Conference*, pp. 156–163, November 2000.

[10] K. Nagao, Y. Shirai, and K. Squire, "Semantic annotation and transcoding: Making web content more accessible," *IEEE MultiMedia*, vol. 8, pp. 69–81, April 2001.

[11] T. Sullivan and R. Matson, "Barriers to use: Usability and content accessibility on the web's most popular sites," *Proceedings of ACM CUU 2000*, pp. 139–144, November 2000.

[12] "Annotation of web content for transcoding." W3C Note, `http://www.w3.org/1999/07/NOTE-annot-19990710`, July 1999.

[13] "Extensible markup language (xml) 1.0 (second edition)." W3C Recommendation, `http://www.w3.org/TR/2000/REC-xml-20001006`, October 2000.

[14] 斎藤 照花, 林 宏紀, 堀内 靖雄, 市川 熹, "グループ分割による視覚障害者の www アクセシビリティ改善手法," 電子情報通信学会 技術研究報告 (WIT00-40), pp. 7–12, March 2001.

[15] "Xml path language (xpath) version 1.0." W3C Recommendation, `http://www.w3.org/TR/1999/REC-xpath-19991116`, November 1999.