自 然 言 語 処 理 154－20
（２００３．３．７）

HTML形式の表構造の内容解析手法とその応用に関する研究

大谷 貴志　獅々堀 正幹　柘植 覚　北 研二

徳島大学大学院 工学研究科 知能情報工専攻

〒 770-8506 徳島市南常三島町 2-1

e-mail: {takasi, bori, tsuge, kita}@is.tokushima-u.ac.jp

あらまし　WWW 空間上の HTML 文書には，形式的な情報を分かり易く表示するために表が頻繁に掲載されている．これら表構造内には，各項目の上位概念となる属性名や各項目間の関係など，言語学的にも非常に有益な情報を含んでいる．しかし，これらの情報を表構造内から獲得するためには，表内においてどの項目が属性なのか，また，その属性と属性値の関係は行列どちらの方向なのかといった各項目の意味的な関係を解析する技術，すなわち，表の内容解析を行う必要がある．

そこで本稿では，WWW 空間上の表構造から言語的に有用な知識を獲得するために，HTML 形式の表構造に対する内容解析を行う手法を提案する．本手法は，各項目の行列方向に存在する項目群をその項目の文脈として捉える．そして，表内の各項目に意味情報が人手で付与された正解データを学習データとして用い．学習データと解析データでの文脈の類似性に基づいて各項目の意味情報の特定を行う．実際に WWW 上に存在する 300 件の表データを用いた実験の結果，表内各項目の意味情報の特定精度（平均適合率）は 0.92 となり，本手法の有効性を確認した．更に，表内容解析結果を応用した Web アプリケーションとして，問い合わせシステムと読み上げシステムについて述べる．

キーワード　表内容解析，Web アプリケーション，問い合わせシステム，読み上げシステム

A Mehod for Analysis of Table Contents of HTML Format and Its Application

Takashi Otani　Masami Shishibori　Satoru Tsuge　Kenji Kita

Department of Information Science & Intelligent Systems
Faculty of Engineering, Tokushima University

2-1, Minami-josanjima, Tokushima, 770-8506

e-mail: {takasi, bori, tsuge, kita}@is.tokushima-u.ac.jp

Abstract　HTML documents in the WWW space frequently include the table structure, which has a very useful information, such as the meanings and relations of words in the table. In order to extract those information from table structures, we have to specify attribute items and relations between attributes and values in the table. This process is called the tables contents analysis. In this paper, we propose the method to analysis of table contents of HTML format.

From the experiment result using 300 HTML table structures, which are collected from WWW space by hand, it was found that this method can obtain 92 percent as the average precision. Moreover, We also mention the inquiry system and the home page reading system, which are web applications adapting the acquired linguistic knowledge.

key words　table contents analysis, Web application, inquiry system, read out system

1　はじめに

近年のインターネット技術の発展は目覚しく，WWW 空間上には膨大な数の情報が蓄積されるようになった．WWW 空間上に存在する HTML 文書には，構造的な情報を視覚的にも分かり易く伝達するために，表形式の情報が頻繁に掲載されている．特に広告ページには，各商品の価格，発売日，スペックなどを示す数多くの表が掲載されている [1]．これらの表構造内には，行列方向の項目間に関係があり，また，各行と列毎に違った意味を持っている．そして，その意味情報は各行列の最上位の項目から判定可能である [2]．このように，各項目の上位概念となる属性名や各項目間の関係など，言語学的にも非常に有益な情報が含まれているにも関わらず，ネットサーチ・エンジンに代表される従来の WWW アプリケーションでは，表内の情報が殆ど扱われていなかった．

そこで本研究では，HTML 形式の表構造から言語的な知識を獲得することを目的とする．本目的を実現するためには，HTML 形式の表構造に対して，表内の各項目間の対応関係，また各項目が属性もしくは属性値どちらの意味的役割を有するかの識別といった表の内容解析を行う必要がある．本稿では，この表内容解析手法を提案し，取得した言語学的知識の Web アプリケーションへの応用について述べる．

本手法は，まず各ページのどの部分に表が記載されているかを認識するために表の識別を行う．HTML によりブラウザに表を表示するためには，TABLE タグが使用される．しかし，TABLE タグは表を表示するといった目的以外にもページのレイアウトとして頻繁に用いられ，TABLE タグが必ずしも表形式を表しているとは限らない．そのため，TABLE タグによって記載されている部分が表形式を表しているのか，もしくは，レイアウトを表しているのかを判別する表形式の識別処理を行う．

次に，識別処理により表形式と認識されたデータに対してのみ，表の構造を解析し，各項目間の位置関係を求める．この表構造解析処理では，各項目毎にその位置情報（各項目が表内のどの位置に記載されていたかを示す）が得られ，各項目と位置情報の組み合わせがデータベース化される．このデータベースを用いると，入力された項目が表のどの位置に記載されていたかといった情報だけでなく，その項目の行列方向に存在する項目群を高速に検索することが可能になる．また，学習用データに対しては，表構造解析を行った後，各項目に対する属性，属性値の区別といった意味情報を人手で付与したデータベース (表構造データベースと呼ぶ) を作成する．この表構造データベースを用いると，属性もしくは属性値の意味毎に，学習データ内における各項目の行列列方向に存在する項目群を参照することができる．

最後に，学習用データから作成された表構造データベースを参照し，解析対象となる表の内容解析を行う．ここでは，表内の各項目が有する意味情報の違いが，その項目の行列方向に存在する項目群の内容に反映される点に着目した．そして，各項目の意味情報と，行列方向の項目群との関係を判定するための教師データとして表構造データベースを用い，表内容解析を行っている．

以下，**2** において，表から取得可能な言語学的情報について説明し，**3** において，言語学的情報を獲得するための表内容解析法について述べる．その後，内容解析法手法の有効性を検証するため，**4** において表内容解析実験を行い，結果の考察を行う．**5** では，表内容解析結果の応用例として，問い合わせシステムと読み上げシステムについて述べる．最後に，**6** において，本稿のまとめと今後の課題について述べる．

2　表内容解析の必要性

一般に，表は形式的な情報を容易に伝達するために，様々なところで使用される．これら表構造内には，行列方向の項目間に関係があり，また，各行や列に違った意味を持つ．そして，各項目がもつ意味情報は，各行もしくは列の最上位の項目から判定できる．

例えば，図 1 に示す表は，パソコンソフトの価格情報を記載した表[1]であるが，パソコンに詳しくない人でも “ウイルスバリア” という文字列は，最上位の項目を見れば，“製品名” であることが分かる．また，項目間の関係を見れば，このソフトの最安値は “6398 円” であることも分かる．

このように，“製品名” などの意味情報を記載している項目を属性と呼び，属性以外の項目を属性値と呼ぶと，属性と属性値の関係といった言語学的知識が表構造内に形式的に記述されている．そ

[1] http://www.kakaku.com/sku/price/soft.htm

のため，一般的な自然言語文に比べて表構造内からの方が，言語学的知識を獲得しやすいと考えられる．

しかし，HTML ファイルを見ただけでは，どの項目が属性なのか，判断できない．よって，表構造内に含まれる有益な情報を得るためには，どの項目が属性なのか，どの項目が属性値なのかといった，表内容解析をする必要がある．

図 1: パソコン価格表の例

3 表内容解析

3.1 全体概要

表内容解析手法の概要図を図 2 に示す．本手法は，学習モジュールと解析モジュールの二つに分かれる．学習モジュールでは，学習用表データ内の各項目が有する意味情報をデータベース化する．解析モジュールでは，解析用表データ内の意味情報が未知な各項目に対して，データベース化された情報を参照することで，意味情報を決定する．

まず，両モジュールで用いられる表形式の識別処理と表構造解析処理について説明する．表形式の識別処理では，入れ子になっている TABLE タグを最小単位に分割する．TABLE タグは表を表す他に，ページのレイアウトとしても用いられるので，分割処理の後，各 TABLE タグが表形式を表しているのかレイアウトとして用いられているのかを判別する．次に表構造解析処理では，表形式と判定されたデータに対して，表内に存在する各項目毎に位置情報を求める．ここで，位置情報は 0 と 1 のコンパクトなビット列によって表現されており，各項目間の行列方向の関係を高速に検索することが可能である [3]．なお，表形式の識別処理については **3.2** ，表構造解析処理については **3.3** で詳細を述べる．

学習モジュールの意味情報付与処理では，表構造解析で得られた各項目に対して，人手により意味情報（各項目の属性，属性値の区別，属性関係が有効な行列方向の向き等）を付与し，表構造データベースに蓄える．つまり，この表構造データベース内には，各項目と行列方向に存在する上位下位項目群との関係（この関係を表構造内での「文脈」と呼ぶ）が格納されている．また，各項目には意味情報が付与されているので，学習用データ内で各項目が有する意味に対応した文脈情報を表構造データベースから検索することが可能になる．この意味情報の付与については **3.4** で詳細を述べる．

一方，解析モジュールの表内容解析処理では，解析対象の表内各項目が有する文脈と表構造データベース内に格納されている文脈とを比較し，各項目の意味情報を決定する．なお，文脈間の相互情報量 [4] を用いて確率的に計算する．アルゴリズムの詳細は **3.5** で示す．

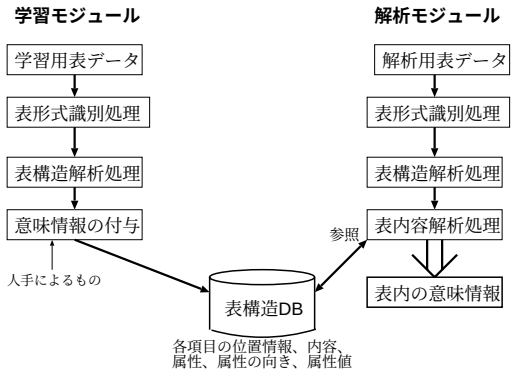


図 2: 本手法の概要図

3.2 表形式の識別処理

HTML の TABLE タグでは，TABLE タグで囲まれた内部に再度 TABLE タグを使用することが可能である．そのため，TABLE タグが入れ子状態で使用されているページが頻繁に存在する．そこで，まず表形式の識別処理では TABLE タグを図 3に示すように最小限に分割する．

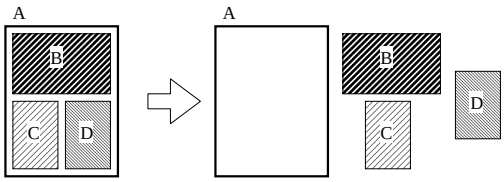


図 3: TABLE タグの分解の例

更に，TABLE タグは表形式を表す他にレイアウト目的で用いられているページが多い [5]．そこで，表形式を表示するための目的だけで使用されている TABLE タグを特定する．本手法では，表内の行数，列数，セル内文字列数によって判別する．判別基準は以下のとおりである．

- 1 行 n 列，n 行 1 列の TABLE タグはレイアウトである
- セル内の文字列数が 0 の場合，TABLE タグはレイアウトである

図 4 のページ[2]に対して行った，表形式の識別処理の実行例を図 5 に示す．最小限に分割された TABLE タグは，1 つずつファイルに保存され，そのファイル名が左のフレームに表示される．このとき，TABLE タグが表形式なのか，レイアウトなのかを判別し，ファイル名の隣に表示される．そして，ファイル名をクリックすると，そのファイルに保存された表が表示される．

図 4: 表形式の識別処理を実行するページの例

[2]http://www.sun-inet.or.jp/ ysannmei/reizouko.htm

図 5: 表形式の識別処理の実行例

3.3 表構造解析処理

表を解析し，位置情報を生成する方法について，図 6 の表を例として説明する．まず，表内には多数の分割軸が存在するが，各分割軸の視点をカッティングポイントと呼ぶ．図 6 の表では，1，9，11 が縦方向の分割軸に対するカッティングポイント，2，4，6，8 が横方向のものである．

本手法では，2 つの制約事項に従って，適用すべきカッティングポイントの順番を制御し，縦→横の順に表を分割しながらビット列を生成する．まず，最初の制約事項を以下に示す．

- **制約事項 1**

  より外側で，原点よりに位置するカッティングポイントから分割を行う．

より外側から分割するのは，複数の小さな表が集まって複雑な表を形成していると考え，できる限り大きな表に分割しながら，表を細分かするためである．また，原点よりから順次分割するのは，同じ行 (列) に属する項目の出現順を正しく位置情報に反映させるためである．図 6 (a) の表において，縦方向のカッティングポイントは，1 → 9 → 11 という順番で選択され，横方向は，2 → 4 → 8 → 6 ではなく，2 → 4 → 6 → 8 となる．つまり，1 の後に 2 で分割された結果，図 6 (b) のように，項目 A の位置情報 (“00”) と項目 B の位置情報 (“10”) が確定する．

次に，2 番目の制約事項を以下に示す．

- **制約事項 2**

適切なカッティングポイントが存在しない場合は，位置情報が未確定な全ての領域にダミービット値'1' を与える．

尚，上記の制約で,「適切でない」とされるカッティングポイントは，以下の条件を満たすものである．

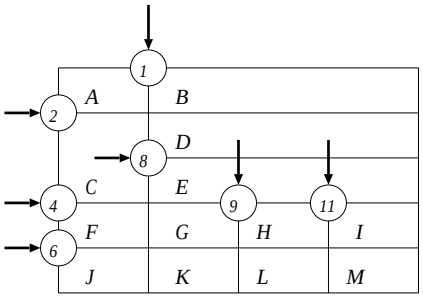
- **条件 1**

  カッティングポイントの上位に，位置情報が未確定の領域が存在する．

- **条件 2**

  カッティングポイントが存在しない．

この制約条件に従って縦→横の順で表を分割し，ビット列を求めると，図 6 (b) のようなビット列が求められる．このように位置情報をビット列で表すことで，ビット列を操作することによりその項目の行・列方向における上位項目を高速に検索することができる．



(a)表構造の例

00	10		
0110	11101010		
	11101011		
011110	111110110	11111011110	11111011111
011111	111111110	11111111110	11111111111

(b)位置情報生成例

図 6: 表構造と位置情報の生成例

この処理部で生成された位置情報 (ビット列) は，奇数ビットが行 (横) 方向，偶数ビットは列 (縦) 方向の位置関係を表す．この縦・横の位置情報を表すビット列を操作することにより，各関係を持つ項目が容易に検索できる．

例えば，図 6(a) の表で項目 A の縦方向の下位項目を知りたい場合，まず，項目 A の位置情報 (“00”) を得た後，偶数ビットの第 2 ビット目の'0' を下位を表すビット値 '1' に反転する．そして，“01” を接頭辞に持つ位置情報を検索することにより，項目 C，F，J，が得られる．横方向の下位項目を知りたい場合，奇数ビットの '0' を '1' に反転することによって，“10” を接頭辞に持つ項目 B が検索される．

また，横方向の上位項目を知りたい場合奇数ビットの '1' を '0' に，縦方向の上位項目を知りたい場合偶数ビットの '1' を '0' に反転することによって，検索できる．

3.4 意味情報の付与

意味情報の付与では，表構造解析した各項目が属性であるか，属性値であるか，また，属性の方向はどちらかを，人手によって付与する．図 7 に意味情報付与の例，図 8 に意味情報付与の実行例を示す．

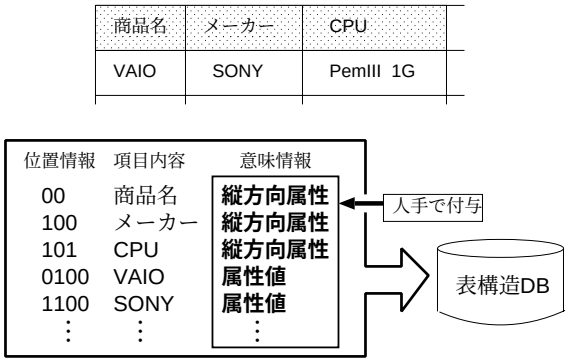


図 7: 意味情報付与の例

図 8: 意味情報付与の実行例

図 8 は，図 5 の右フレームの表を表構造解析し

た結果である．四角で囲った部分を例に挙げると，“ 00 ” という位置情報を持った “型名” という文字列は,「縦方向に意味関係を持った属性」なのか,「横方向に意味関係を持った属性」なのか,「縦横両方向に意味関係を持った属性」なのか，または「属性値」なのかをラジオボックスで選択する．この処理部では，このように，表の各項目毎に意味情報を人手で付与し，表構造データベースに登録する．

3.5 表内容解析処理

表内容解析では，データベースを参照し，表の各項目が属性であるか属性値であるかを判断する．その手法を，図 9 の n 行 m 列の表を使って示す．ただし，初期状態として表内 1 行 1 列目の項目を項目 x とする．

商品名	メーカー	CPU
VAIO	SONY	PemIII 1G
Mebius	SHARP	PemIII 1.5G

図 9: 内容解析する表の例

- **手順 1**

  項目 x (商品名) を属性と仮定したときの下位項目 y_k (VAIO，Mebius，$\cdots$) との相互情報量 I_{attri} と，項目 x を属性値と仮定したときの項目 y_k との相互情報量 I_{value} を求める．

$$I_{attri}(x,y) = \sum_{k=2}^{n} \log \frac{P_{attri}(x,y_k)}{P_{attri}(x)P(y_k)} \qquad (1)$$

$$I_{value}(x,y) = \sum_{k=2}^{n} \log \frac{P_{value}(x,y_k)}{P_{value}(x)P(y_k)} \qquad (2)$$

ここで，$P_{attri}(x)$ は項目 x (商品名) が属性としてデータベースに登録されている確率，$P(y_k)$ は項目 y_k (SONY，SHARP，$\cdots$) が属性値として登録されている確率，$P_{attri}(x,y_k)$ は項目 x が属性であり，かつ，その下位項目に項目 y_k が登録されている確率，$P_{value}(x)$ は項目 x (商品名) が属性値として登録されている確率，$P_{value}(x,y_k)$ は項目 x が属性値であり，かつ，その下位項目に項目 y_k が登録されている確率を表す．

- **手順 2**

  手順 1 で求めた I_{attri} と I_{value} の大きさを比較し，I_{attri} の方が大きければ項目 x は属性，I_{value} の方が大きければ項目 x は属性値とする．

- **手順 3**

  メーカー， CPU ，$\cdots$，と一行目の項目全てに対し，手順 1 と手順 2 を繰り返し，属性値と判断される項目より，属性と判断される項目のほうが多ければ， 1 行目は属性と判断される．

- **手順 4**

  一列目の VAIO，Mebius，$\cdots$ も同様に相互情報量を比べ， 1 列目が属性かどうかを判断する．

- **手順 5**

  表内容解析終了．

以上の手順によって，図 10 のような 1 行目が属性である表と， 1 列目が属性である表， 1 行目も 1 列目も属性である表の 3 種類の表のいずれかと判断される．

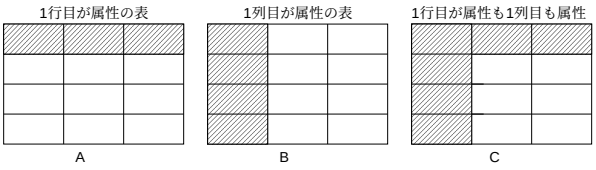


図 10: 解析される表のタイプ

4 表内容解析実験

本手法を用いた表内容解析の有効性を検証するため，パソコンに関する表，車に関する表，家電製品に関する表を各 100 件ずつ集め，表の各項目が属性を表しているのか，属性値を表しているのかを解析する実験を行った．

4.1 実験条件

実験に用いる表データは，手作業で集めたパソコンに関する表 100 件，車に関する表 100 件，家

電製品に関する表 100 件，合計 300 件である．実験は，各種の表別で行う．まず，実験に用いる 100 件のデータを 10 件ずつランダムに 10 分割する．そして初めの 10 件を解析データとし，残りの 90 件を学習データとする．次に，先程と違う 10 件を解析データとし，残りの 90 件を学習データとする．この操作を 10 回繰り返すことによって，100 件全ての表に対して実験を行う．実験結果の評価基準としては，解析データに対する表内容解析結果を各項目毎に正解データと比較し，その結果を平均適合率で評価した．

4.2 実験結果

パソコンに関する表，車に関する表，家電製品に関する表，各 100 件ずつに対して表内容解析実験をした結果，平均適合率は表 1 のような結果になった．

表 1: 実験結果

表の種類	平均適合率
パソコン	0.95
車	0.87
家電製品	0.95

解析できなかったものに，図 10 の 3 種類以外の表，例えば図 11 のような，1 行目と同じ内容が下位の行に存在し，その行も属性となる表が挙げられる．本手法では 1 行目と 1 列目しか，属性か属性値かを解析していないのが原因だと考えられる．そこで，全項目に対して解析をすることで解決できるのではないかと考えている．

また,「 APPLE 」や「 HDD 」など，属性とも属性値ともなり得る項目に対しての誤りが 21 件ほどあった．本手法では，相互情報量を比べるのに，(1) 式 (2) 式では $k = 2$ から $k = n$ までの和を比べていたが，k 毎に比べることにより改善されるのではないかと考えている．

図 11: 1 行目以外の行にも属性がある表の例

5 表内容解析結果の応用

前述したように， WWW 空間上の表構造には非常に多くの有益な情報が含まれている．そして，表内容解析をすることによって様々な分野に応用することができる．ここでは，表内容解析の結果を応用したシステムについて述べる．

5.1 問い合わせシステム

一般に，ネットサーチエンジンでは,「組織名」,「人名」,「専門用語」などの固有名詞が検索語として指定されることが多い．しかし固有名詞には，同じ表記でも違った意味 (意味的多義性) を持つものが多く存在する [6]．そのため，固有名詞が持つ意味的多義性に気づかずに検索すると，従来のネットサーチエンジンでは，検索結果に多くのノイズが含まれ，検索精度を低下させる原因となることがある．例えば,「大塚」という固有名詞を検索語として指定すると,「組織名」の意味を持つ「大塚」と「人名」の意味を持つ「大塚」,「地名」の意味を持つ「大塚」が検索されてしまう．このような固有名詞が持つ問題点に対し，我々は，表構造内に存在する固有名詞の意味情報が，各行・列の上位方向の項目 (属性) から判定可能と考え，固有名詞の意味情報を考慮した問い合わせシステムを考えている．

5.2 読み上げシステム

視覚障害者のために，ブラウザに表示されている内容を音声によって読み上げるシステム [7] があ

る．しかし，従来の読み上げシステムでは，タグを除去して読み上げるだけなので，表に空欄があると各項目間の関係にずれが生じるといった問題点が挙げられる．例えば，図 12 をホームページリーダーで読み上げると,「メーカー，車種，売価，走行距離，保険，年式，車体色，アプリリア，00 年アプリリアエリア 518BK，319000 円，Km，′00，ワークスカラー，···」と読み上げられる．これでは視覚障害者は保険の項目が空欄であることは分かりにくい．そこで，表内容解析を行うことにより,「メーカーはアプリリア，車種は 00 年アプリリアエリア 518BK，売価は 319000 円，···」と，分かりやすく読み上げることができると考えている．

メーカー	車種	売価	走行距離	保険	年式	車体色
アプリリア	00年アプリリアエリア 518BK	319,000 円	Km		'00	ワークスカラー
アプリリア	アプリリア (ON)RSV1000ミッレ	1,580,000 円	Km	検付		黒
ホンダ	NSR250	139,000 円	15,975 Km		'87	白/赤
ホンダ	NSR250-2	230,000 円	Km			白/赤
ホンダ	スティード400	420,000 円	Km	検受		黒

図 12: 車に関する表

6 まとめ

本稿では，HTML 形式の表構造の表内容解析手法を提案した．そして，本手法の有効性を検証するために，表内容解析実験を行った．実験には，パソコンに関する表と，車に関する表，家電製品に関する表を，各 100 件ずつ用いた．その結果，平均適合率は 0.92 となり，本手法は有効であることが確認された．

また，表内容解析を応用したシステムである，問い合わせシステムと，読み上げシステムについて述べた．

今後の課題として，様々な表での実験，解析精度の向上，応用システムの実装などが挙げられる．

参考文献

[1] 井上香織，高橋克巳，“検索のための広告文書構造化”，情報処理学会第 57 回全国大会論文集，1V–4, pp. 207–208, 1998.

[2] 吉田稔，鳥澤健太郎，辻井潤一，“表形式からの情報抽出手法”，言語処理学会第 6 回年次大会，pp. 252–255, 2000.

[3] 獅々堀正幹，岩口義広，青江順一，“HTML 形式の表構造に対する一索引手法”，情報処理学会データベースシステム研究会資料，125–40, pp. 305–312, 2001.

[4] 北研二，中村哲，永田昌明，“音声言語処理”，森北出版株式会社, 1996.

[5] 杉原勇，“WWW 上に存在する表構造データの分析と抽出に関する研究”，徳島大学卒業論文, 2001.

[6] 西野文人，落合亮，“抽出情報の実体あいまい性の解消”，言語処理学会第 6 回年次大会ワークショップ論文集，pp. 41–48, 2000.

[7] ホームページリーダー homepage. http://www-6.ibm.com/jp/accessibility/soft/hpr.html