RN-002

ユーザの語彙力に適応した読みを付与する Web 読解支援システム
Web Based Reading Support System Adapting to the Vocabulary Levels of Individual Users

溝渕 昭二† 安藤 一秋‡
Shoji Mizobuchi Kazuaki Ando

1. はじめに

近年，情報通信技術の発展や情報メディアのデジタル化に伴い，様々な立場の人々が World Wide Web（以降では，Web と記す）上にある Web ページを読解するようになっている．

Web ページを読解する際の問題点の一つに，未習あるいは想起困難な語句の存在が挙げられる．Web ページの読解中にそれらの語句が頻出すると，その行為に対して様々な悪影響が生じる．例えば，読解にかかる時間が増加したり，読解から得られる知識が減少したり，さらには，読解に対する意欲が低下したりするなどである．これらの影響は，すべてのユーザに起こりうるが，特に語彙力が未発達なユーザ（小中高の児童生徒，非母国語の学習者，特定分野の初学者など）については，それらの影響が深刻化する．

語彙力が未発達なユーザに対する支援の一つに，Web ページ内の語句に読みを振る方法がある．実際，この方法によりサービスを提供している Web サイト（Yahoo きっずや YOMOYOMO など）も存在する．

しかしながら，既存の Web サイトでは，ユーザの語彙力は静的なものとして扱われ，ユーザから申告された語彙力はずっと変わらぬまま利用される．したがって，申告された語彙力と真の語彙力との間に隔たりがある場合，読みが適切に振られず，前述したような弊害が生じる．また，継続した利用によりユーザの語彙力が変化しても，それに応じた読みを振ることができない．

そこで，本論文では，Web ページの読解支援を行うことを目的に，項目反応理論（Item Response Theory）[1]を用いて推定されたユーザの語彙力に合わせて，Web ページ内の語句に読みを振るシステムを提案する．

本論文の構成を次に記す．つづく 2 節では関連研究について述べる．次に，3 節では提案システムについて述べる．次に，4 節では提案システムを評価した結果について述べる．最後に，5 節ではまとめと今後の課題について述べる．

2. 関連研究

Web ページ内のテキストに何らかの処理を施して，その読解を支援しようとする研究は多く存在する[2]．近年は，ユーザの目標や能力に適応させて，その処理内容を変動させる手法も登場している[3]．

Web ページ内のテキストに対する処理方法とユーザへの適応能力という二つの観点から関連研究を分類したものを表 1 に示す．

Web ページ内のテキストに対する処理方法は，注釈，置換，要約の 3 種類に大別できる．注釈は，元の内容を改変せずにその理解を促すような情報を付与する方法である．

表 1 関連研究の分類

		適応能力	
		静的	動的
処理方法	注釈	榎本ら[4]	江原ら[5] 光原ら[6]
	置換	榎本ら[7] 松吉ら[8]	
	要約	原[9]	

置換は，ある表現をその意味を保持したまま別の表現に変更する方法である．そして，要約は，ある表現をその真意が伝わるように縮約する方法である．

ユーザへの適応能力は，その能力を静的に捉える方法と，動的に捉える方法の 2 種類に大別できる．静的に捉える方法は，ユーザの能力を全く把握しないか，自己申告された能力を変更することなく継続して利用するものである．動的に捉える方法は，ユーザの能力を把握して，それに適応するように支援を行うものである．

榎本ら[4]は，漢字には複数の読みがあるという点に着目した漢字の読み振りシステムを提案している．これは，自己申告された語彙力に基づいて，Web ページ内の漢字に読みを付与している．したがって，表 1 では注釈と静的に該当する箇所に位置づけられる．

江原ら[5]は，Web ページ内にある英単語に対して和訳を付与するシステムを提案している．また，光原ら[6]は，Web ページの関連知識を自動的に提示するシステムを提案している．これらは，ブラウザの操作履歴から推定したユーザの能力に応じて Web ページ内の語句に和訳や関連知識などの情報を付加している．したがって，表 1 では注釈と動的に該当する箇所に位置づけられる．

榎本ら[7]は，Web ページ内の漢字を読みに置換するシステムを提案している．また，松吉ら[8]は，難易度と文体の制御が可能な日本語機能表現の言い換え手法を提案している． これらは，自己申告された語彙力あるいはパラメータに基づいて，Web ページ内の語句を同じ意味の語句に置換している．したがって，表 1 では置換と静的に該当する箇所に位置づけられる．

原[9]は，子供の嗜好や特性を考慮して，一般の Web ページを子供向けページに要約するブラウザを提案している．子供向けページでは，一般の Web ページに含まれるトピックを海中に漂う泡で，トピックの詳細を絵本で表現している．これは，小学生を対象としているものの，ユーザの能力とは無関係に Web ページの内容を泡と絵本という形式で要約している．したがって，表 1 では要約と静的に該当する箇所に位置づけられる．

本システムは，江原らや光原らと同じく注釈と動的に該当する箇所に位置づけられる．しかし，それらが訳語や関連知識を付加するのに対して，本システムは読みを付加す

† 近畿大学 Kinki University
‡ 香川大学 Kagawa University

る点が異なる．本システムと同じく読みを付加するシステムは，すでに榎本らや既存の Web サービスによって提案されている．しかし，それらはユーザの能力を静的に捉えており，本システムのようにユーザへの適応能力は持たない．

3. 提案システム

本システムは，ユーザの語彙力と文字の難易度に基づいて，そのユーザが閲覧する Web ページ内の単語に読みを付与する．ユーザの語彙力は，それらに付与した読みに対して実際にユーザが行った操作から推定される．一方，文字の難易度は，事前に収集したテスト結果から推定される．これら 2 種類のパラメータの推定には，項目反応理論が利用される．

3.1 構成

本システムは，文字登録部，ページ表示部，操作応答部の三つのコンポーネントから構成される．本システムの構成を図 1 に示す．

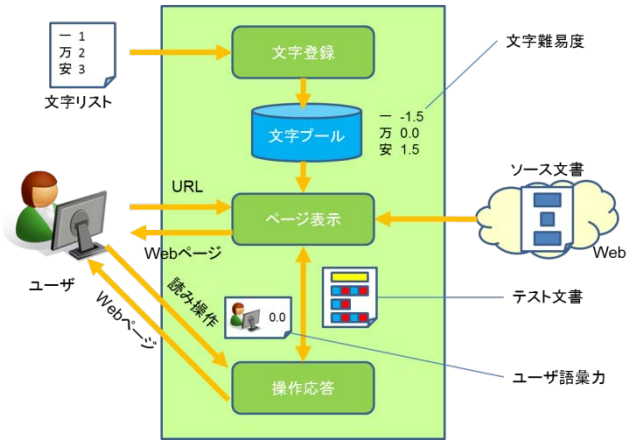

図 1 システム構成

3.1.1 文字登録部

文字登録部は，ランク付けされた文字のリスト（以後，文字リストと記す）を元に，各文字の難易度を推定し，その結果を文字プールに登録する．

本コンポーネントにおける処理の概要を図 2 に示す．また，その手順は以下のとおりである．

1. 文字リストから項目反応データを作成する（3.2.2 節を参照）．
2. 項目反応データから各文字の難易度を推定する．
3. 文字と難易度の組を文字プールに登録する．

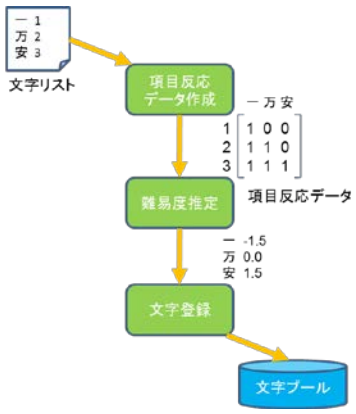

図 2 文字登録部の処理手順

3.1.2 ページ表示部

ページ表示部は，ユーザにより指定された URL にある HTML 文書（以後，ソース文書と記す）に読みを付与する．そして，読みが付与された HTML 文書（以後，テスト文書と記す）を Web ページとして表示する．この際，ユーザの語彙力を文字プールに登録された文字の難易度から推定し，それらを比較することにより，付与した読みを実際に Web ページ内で表示するかどうかを決定する．

本コンポーネントにおける処理の概要を図 3 に示す．また，その手順は以下のとおりである．

1. Web から取得したソース文書を解析し，DOM ツリーを作成する．
2. 形態素解析により，DOM ツリー内のテキストを単語に分割し，それに対して読みを付与する．
3. 単語，ユーザの語彙力，および，文字プール内にある文字の難易度を元に，項目反応データを作成する（3.2.3 節を参照）．
4. 項目反応データからユーザの語彙力を推定する．
5. 読みの切り替え操作を取得するためスクリプトを挿入した DOM ツリーからテスト文書を生成する．
6. テスト文書を Web ページとしてユーザに表示する．

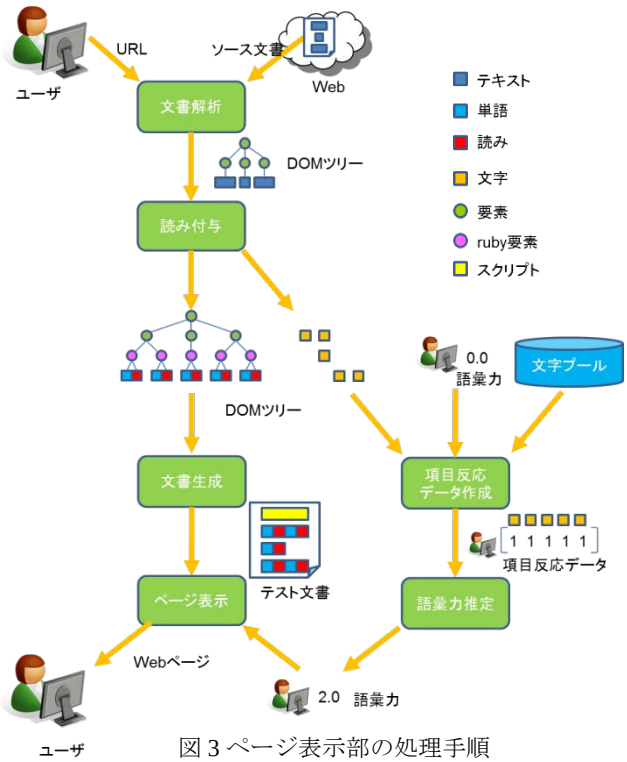

図 3 ページ表示部の処理手順

3.1.3 操作応答部

操作応答部は，ユーザによる読みの切り替え操作を元に，ユーザの語彙力を推定する．そして，それに基づいて Web ページ内の読みの表示を切り替える．

本コンポーネントにおける処理の概要を図 4 に示す．また，その手順は以下のとおりである．

1. ユーザによる読みの切り替え操作を項目反応データに反映させる（3.2.3 節を参照）．
2. 項目反応データからユーザの語彙力を推定する．
3. ユーザの語彙力に応じて，Web ページ内の読みの表示を切り替える．

3.2 項目反応理論を利用したパラメータ推定

項目反応理論（Item Response Theory）[1]は，個人の能力値や項目の難易度などのパラメータを，項目への反応から求めようとする理論である．以降では，項目反応モデルの一つであり，本システムで使用する 1 パラメータロジスティックモデルと，文字の難易度およびユーザの語彙力を推定するのに使用する項目反応データの作成方法について説明する．

3.2.1 1 パラメータロジスティックモデル

1 パラメータロジスティックモデルにおいて，n 人の被験者の能力値を$\theta_i(1 \le i \le n)$，m 個の項目の難易度を$b_j(1 \le j \le m)$，被験者 i の項目 j に対する反応をu_{ij}（正答のとき 1，誤答のとき 0）とするとき，能力値θ_iを持つ被験者 i が項目 j に正答する確率は，式（1）で表わされる．

$$P_{ij} = P(u_{ij}=1|\theta_i, b_j) = \frac{1}{1+exp(\theta_i + b_j)} \quad (1)$$

そして，m 個の項目からなるテストに対する被験者 i の項目反応パターンが，$\boldsymbol{u_i}$=$(u_{i1},\cdots,u_{im})$になる確率は，式（2）で表わされる．

$$P(\boldsymbol{u_i}|\theta_i, b_1, \cdots, b_m) = \prod_{j=1}^{J} P_{ij}{}^{u_{ij}}(1-P_{ij})^{(1-u_{ij})} \quad (2)$$

さらに，n 人の被験者の項目反応パターンからなる項目反応データが，$\boldsymbol{U}$=$(\boldsymbol{u_1},\cdots,\boldsymbol{u_n})^{\mathrm{T}}$になる確率は，式(3)で表わされる．

$$P(\boldsymbol{U}|\theta_1,\cdots,\theta_n, b_1,\cdots,b_m) = \prod_{i=1}^{n}\prod_{j=1}^{m} P_{ij}{}^{u_{ij}}(1-P_{ij})^{(1-u_{ij})} \quad (3)$$

被験者の項目反応データは，テストの結果として得られるので，それを使って式(3)を最尤推定すれば，被験者の能力値と項目の難易度が推定できる．

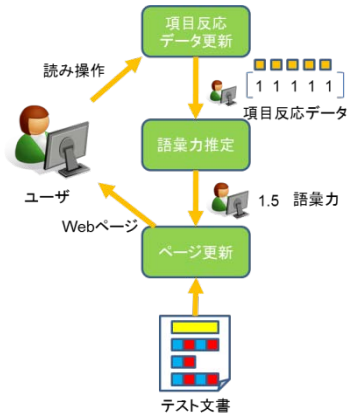

図 4 操作応答部の処理手順

3.2.2 文字難易度の推定

文字の集合をC，ランクの集合をR，文字リスト内の文字とランクの組からなる集合をCRとするとき，各文字の難易度を推定するのに使用する項目反応データは，$|R| \times |C|$型の行列$U = (u_{ij})$で表わされ，その要素は式(4)のとおりに定義される．

$$u_{ij} = \begin{cases} 1, & (c_j, r) \in CR \wedge r_i \ge r \\ 0, & \text{上記以外} \end{cases} \quad c_j \in C, r, r_i \in R \quad (4)$$

3.2.3 ユーザ語彙力の推定

ソース文書内に登場する文字$c_j \in C$の難易度を$d(c_j)$，更新前のユーザの語彙力をθとするとき，更新後のユーザ語彙力を推定するのに使用する項目反応データは，$1 \times |C|$型の行列$U = (u_{ij})$で表わされ，その要素は式(5)のとおりに定義される．

$$u_{ij} = \begin{cases} 1, & d(c_j) \le \theta \\ 0, & \text{上記以外} \end{cases} \quad (5)$$

ユーザが切り替え操作を行う単語を$w = c_{s_1}, c_{s_2}, \cdots, c_{s_l}$とするとき，この難易度$d(w)$は式(6)のとおりに定義される．

$$d(w) = \max \mathrm{D} \quad D = \{d(c_j) | d(c_j) \le \theta\} \quad j = s_1, \cdots, s_l \quad (6)$$

ユーザが単語$w = c_{s_1}, c_{s_2}, \cdots, c_{s_l}$の読みを表示する場合，項目反応データ$U$の要素を式(7)のとおりに更新する．

$$u_{is_k} = \begin{cases} 0 & d(c_{s_k}) > d(w) \\ u_{is_k} & \text{上記以外} \end{cases} \quad (7)$$

ユーザが単語$w = c_{s_1}, c_{s_2}, \cdots, c_{s_l}$の読みを消去する場合，項目反応データ$U$の要素を式(8)のとおりに更新する．

$$u_{is_k} = \begin{cases} 1 & d(c_{s_k}) > d(w) \\ u_{is_k} & \text{上記以外} \end{cases} \quad (8)$$

3.3 実装

本システムは，Web アプリケーションであり，クライアントサイドは JavaScript，サーバサイドでは Java を使って実装している．

本システムの実行画面を図 5 から図 8 までに示す．

図 5 は，本システムのトップページである．本システムを利用するユーザは，まずこのページにて最初に閲覧するページの URL と自身のランクを入力する．

図 6 は，ユーザが入力した URL にある Web ページが本システムにより表示されたものである．

図 7 は，読みの表示を切り替えるために本システムが Web ページに埋め込んだスイッチである．このスイッチは，読みを振った語句をマウスでポイントすると表示される．このスイッチをクリックすることで，該当する語句の読みを表示したり，消去したりできる．

図 8 は，スイッチの切り替え操作を何回か行った後の Web ページの内容である．オレンジの線より上側にある語句に対して切り替え操作を行った結果，それより下側にある語句のうち，ユーザの語彙力より難易度が高い語句に対して自動的に読みが振られている．

図 5 トップページ

図 6 本システムにより表示された Web ページ

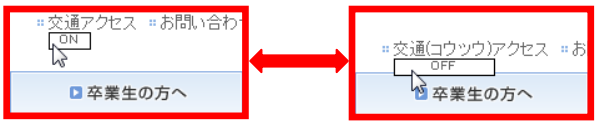

図 7 読み切り替えスイッチ

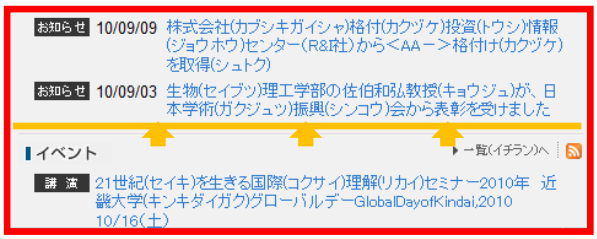

図 8 スイッチの切り替え操作を何回か行った後の Web ページの内容

4. 評価

本システムを評価するために，シミュレーションを実施した．以降では，その方法と結果について述べる．

4.1 方法

シミュレーションでは，まず，人手で作成した文字リストを本システムに登録して，それに含まれる各文字の難易度と，各ランクに属するユーザの語彙力を推定した．次に，各ランクに属するユーザの行動を模倣する Web クライアント（以降では，仮想ユーザと記す）に本システムを操作させた．つまり，所与の URL にあるソース文書にアクセスする操作と，それにより得たテスト文書内にある読みの表示を切り替える操作を行わせた．

本シミュレーションで使用する文字リストには，文部科学省の「小学校学習指導要領付録学年別漢字配当表」に記載された漢字とその配当学年を収録した．表 2 に文字リストの概要を示す．

文字リストの登録時に推定された漢字の難易度と各ランクに属するユーザの語彙力をそれぞれ，表 3 と表 4 に示す．

本シミュレーションにおいて，仮想ユーザがアクセスするソース文書の URL を表 5 に示す．

表 2 文字リストの概要

ランク（学年）	個数	実例
1	80	一，字，森
2	160	刀，科，曜
3	200	丁，屋，題
4	200	士，案，議
5	185	久，益，護
6	181	干，株，臓
合計	1006	

表 3 ランク別推定漢字難易度

学年	難易度
1	−16. 58
2	−10. 03
3	−4. 23
4	1. 23
5	6. 75
6	13. 10

表 4 ランク別推定ユーザ語彙力

学年	語彙力
1	−13. 66
2	−7. 25
3	−1. 50
4	4. 03
5	9. 94
6	21. 30

表 5 ソース文書の URL

番号	URL
1	http://kids.yahoo.co.jp/
2	http://www.aozora.gr.jp/cards/000879/files/127_15260.html
3	http://www.sansu.org/
4	http://www.unesco.or.jp/contents/isan/about.html
5	http://sc-smn.jst.go.jp/
6	http://www.nou-taiken.net/
7	http://www.jaxa.jp/
8	http://www.asahi.com/
9	http://ja.wikipedia.org/wiki/%E9%A2%A8%E5%8A%9B%E7%99%BA%E9%9B%BB
10	http://ja.wikipedia.org/wiki/%E3%82%AB%E3%83%96%E3%83%88%E3%83%A0%E3%82%B7

表 6 未知状態からの語彙力推定結果

ランク	1	2	3	4	5	6
開始推定語彙力	16.19	16.19	16.19	16.19	16.20	16.20
完了推定語彙力	-13.30	-6.98	-1.19	3.86	9.77	16.20
事前推定語彙力	-13.66	-7.25	-1.50	4.03	9.94	21.30
収束率	0.52	0.49	0.45	0.43	0.47	0.00

表 7 1 ランク下の語彙力と推定されている状態からの語彙力推定結果

ランク	2	3	4	5	6
開始推定語彙力	-13.30	-6.98	-1.19	3.86	9.77
完了推定語彙力	-6.98	-1.19	3.86	9.77	16.20
事前推定語彙力	-7.25	-1.50	4.03	9.94	21.30
収束率	0.53	0.50	0.49	0.51	0.58

表 8 1 ランク上の語彙力と推定されている状態からの語彙力推定結果

ランク	1	2	3	4	5
開始推定語彙力	-6.99	-1.19	3.86	9.77	16.20
完了推定語彙力	-13.30	-6.98	-1.19	3.86	9.77
事前推定語彙力	-13.66	-7.25	-1.50	4.03	9.94
収束率	0.57	0.53	0.45	0.47	0.47

仮想ユーザが読みの表示を切り替える操作は，文字リストの登録時に推定されたユーザの語彙力（以降では，事前推定語彙力と記す）に基づいて，次のとおりに行った．

- 事前推定語彙力より高い難易度の単語に読みが振られていない場合，読みを付加する操作を実施
- 事前推定語彙力より低い難易度の単語に読みが振られている場合，読みを消去する操作を実施
- 上記以外の場合．何もしない

4.2 結果

未知の状態からの語彙力の推定結果を表 6 に，ユーザの語彙力を 1 ランク下と推定している状態からの語彙力の推定結果を表 7 に，ユーザの語彙力を 1 ランク上と推定している状態からの語彙力の推定結果を表 8 に示す．

各表にあるランクは，仮想ユーザのランクである．開始推定語彙力と完了推定語彙力は，それぞれ，テスト文書の取得時と読み操作の完了時に本システムにより推定されたユーザの語彙力である．収束率は，仮想ユーザの語彙力より難易度の高い異なり単語数のうち，完了推定語彙力に収束するまでに操作した異なり単語数の比率である．

これらの結果から，ユーザの語彙力を推定するには，Web ページ内にあるユーザの語彙力より高い難易度の単語の約半数に対して読みの操作を行うと，ほぼ正確にユーザの語彙力の推定できることが判明した．

ページ表示部と操作応答部の処理時間を表 9 に示す．また，ソース文書，テスト文書，読み操作の応答のデータサイズを表 10 に示す．

ページ表示部において生成されるテスト文書は，元のソース文書に読み等を付与したものである．したがって，テスト文書のデータ量は，元のソース文書と比較して多くなる．今回のシミュレーションでは，その倍率は平均で 3.3 倍となった．また，テスト文書は，ソース文書を加工したり，ユーザの語彙力を推定したりする処理を経て生成される．したがって，仮想ユーザがテスト文書を取得するのに要する時間は，直接ソース文書を取得するのと比較して長

表 9 処理時間（秒）

URL	ソース文書	テスト文書	倍率	読み操作
1	0.696	1.958	2.8	0.012
2	0.159	3.108	19.5	0.015
3	0.354	1.320	3.7	0.013
4	0.130	1.270	9.8	0.012
5	3.668	5.116	1.4	0.010
6	0.171	0.925	5.4	0.011
7	1.932	3.108	1.6	0.009
8	0.343	2.819	8.2	0.012
9	6.087	12.768	2.1	0.013
10	5.498	10.570	1.9	0.012
Avg.	1.904	4.296	5.6	0.012

表 10 データサイズ（バイト）

URL	ソース文書	テスト文書	倍率	読み操作
1	61953	103798	1.7	14
2	21376	107731	5.0	14
3	23978	55977	2.3	14
4	17278	46353	2.7	14
5	42532	77604	1.8	14
6	17693	34467	1.9	14
7	24546	68418	2.8	14
8	70439	153608	2.2	14
9	186991	608559	3.3	14
10	99541	331203	3.3	14
Avg.	56633	158772	2.7	14

くなる．今回のシミュレーションでは，その倍率は，平均で 5.4 倍となった．

操作応答部では，ユーザの読みの切り替え操作に応じて推定されたユーザの語彙力がレスポンスとして返される．そのデータ量は十数バイトであり，システムの処理速度への影響はほとんどない．また，レスポンスを取得するまでに要した時間も平均で 12 ミリ秒と高速である．

5. おわりに

本論文では，ユーザの語彙力に合わせて，Web ページ内の語句に読みを振るシステムを提案した．ユーザの語彙力は，Web ページに付与した読みに対して実際にユーザが行った操作と事前に求めた文字の難易度から項目反応理論を利用して推定される．

また，本システムを評価するために仮想ユーザによるシミュレーションを実施した．この結果，Web ページ内の半数の語句に対して未知あるいは既知の応答を行えば，ユーザの語彙力をほぼ正確に推定できることが判明した．

参考文献

[1] Baker, B. F. & Kim, S., Item Response Theory: Parameter Estimation Techniques, Second Edition, CRC Press (2004).
[2] 大村 彰道(監), 秋田 喜美代(編), 久野 雅樹(編), 文章理解の心理学, 北大路書房 (2001).
[3] Brusilovsky, P., “Adaptive Hypermedia”, User Modeling and User-Adapted Interaction, Vol. 11, pp. 87-110 (2001).
[4] 榎本 聡, 室田 真男, 清水 康敬, “漢字かな自動変換機能等を備えたインターネット学習システムの開発”, 電子情報通信学会論文誌, Vol, J83-D-I, No. 3, pp. 384-394 (2000)
[5] 江原 遥, 二宮 崇, 中川 裕志, “Web 文書中の単語クリックログの解析から未知単語を予測する語義注釈システム”, 情報処理学会研究報告, Vol. 2009, No. 3, pp.1-7 (2009)
[6] 光原 弘幸, 越智 洋司, 矢野 米雄, “Web ページに関連知識の解説をリンクする WBL システム”, 電子情報通信学会論文誌, Vol. J86-D-I, No. 1, pp. 29-38 (2003)
[7] 榎本 聡, 室田 真男, 清水 康敬, “「音訓の読み方」と「ふりがな表記」に対応した漢字かな自動変換サーバの開発”, 教育システム情報学会論文誌, Vol. 17, No. 3, pp. 275-284 (2000)
[8] 松吉 俊, 佐藤 理史, “文体と難易度を制御可能な日本語機能表現の言い換え”, 自然言語処理, Vol. 15, No. 2, pp. 75-99 (2008)
[9] 原 隆浩, “バブルブラウザ：子供向け Web ブラウザの取り組み―手探りの研究開始から手応えをつかむまで―”, 情報処理学会誌, Vol. 51, No. 1, pp. 5-8 (2010)