LK-021

有用性評価のためのウェブページ構造の特徴分析法
A Method of Structural Feature Analysis for Evaluation of Web Pages

蓬莱一朗[†] 岩谷幸雄[‡] 阿部亨[§] 木下哲男[§]
Ichiro Hourai Yukio Iwaya Toru Abe Tetsuo Kinoshita

1. はじめに

検索エンジンの高性能化によって，欲しい記述を含むウェブページの選別が可能になってきた．これは，リンク構造の解析やテキストマッチングをはじめとする検索手法の進化によるものが大きい [1]．しかし，ウェブページ探索の効率化を図るためには，必要な記述を含むページを検索できるだけでは十分ではなく，検索されたページの有用性を事前に評価・提示しユーザによる最終的な選択を支援する機構が必要である．これを実現するために，ウェブページの有用性を評価する種々の手法が提案されている．しかしながら，ユーザにとっての有用性を大きく左右する「ページの見え方」(見やすいページ＝有用なページ) については，それを基にした有用性の評価法が十分には検討されていない．

一般に，ユーザの意図に合致する記述 (テキスト，アンカーテキスト，画像等) が想定されている場合，その記述がページ上のある箇所にまとまって記載されていれば，ユーザにとって検索目標の検出や判定が容易となることから，それらの「まとまり」を含むページを有用なものとして評価できる可能性がある．この「まとまり」特性に着目することは，見たものを何らかの「まとまり」として認識しようとする人間の視覚特性 [2, 3] や，ページのレイアウトに関する研究 [4] などから，十分検討に値するアプローチと考えられる．そこで本研究では，ページ上の「まとまり」特性に基づき，ページの有用性を評価する手法について検討を進めている．

HTML タグの類似度を基に意味的なまとまりを抽出する既存の方法 [5] では，タグ列から類似パターンの検出処理を行うが，これは一般に処理コストが掛かるため，大規模なページ群を解析対象にした場合に問題となると考えられる．また [5] では，意味的まとまりを対象としているため，抽出時にテキストの長さや画像の大きさなども比較のパラメタとしているが，本研究はテキスト，アンカーテキスト，画像それぞれの視覚的セグメントを対象とするためパラメタの選択が必要である．そこで本稿では，セグメント決定の前処理として HTML のレイアウトタグに基づく簡素な初期分割法を提案する．次に，視覚に特化したパラメタを選択してセグメントの結合を行い，主観評価実験に基づいて有効性の検証を行った．

2. HTML の分析によるセグメントの決定

2.1 セグメント決定のための初期分割

提案する手法は，ウェブページの構造情報に基づいて分析対象をセグメントに分割し，得られたセグメント相互間の構造的類似性を基に，1つのまとまりに含まれると判断されるセグメント群を結合していくものである．

[†]東北大学大学院情報科学研究科
[‡]東北大学電気通信研究所
[§]東北大学情報シナジーセンター

まずセグメント分割の処理では，HTML のレイアウトを構成するタグに基づき，以下の処理に従ってウェブページをセグメントに初期分割する．

1. HTML タグにより，ページ上の記述をテキスト・アンカーテキスト (A タグ)・画像 (IMG タグ) の3種類の記述に分類する．
2. 表やレイアウトを構成するタグ (TABLE, TD)，前後に改行を伴ってレイアウトを構成するタグ (P, DIV, CENTER, TH, H1〜H6, PRE) により，セグメントに分割する．
3. セグメントが，レイアウト調整用の微小な画像 (1辺 10Pixel 未満) や短いテキスト (「・」など) のみで構成されている場合，それらを削除する．

以上の処理によって図1のページをセグメントに分割した結果を図2に示す．破線がテーブル要素 (TD) やその他レイアウトを構成するタグによるセグメントを，実線はテーブルタグ (TABLE) で囲まれている部分を示している．図2の結果からわかるように，レイアウトを構成するタグを用いた分割の段階では，多くの小セグメントに分割され，不自然なセグメントとなっているものが多く存在している．そこで，相互に隣接し，その構造が類似するセグメント群を1つのセグメントに結合する．

2.2 類似性に基づくセグメントの結合

隣接するセグメント α，β 間の類似度を $x_{\alpha\beta}$，セグメント α，β 内に含まれるタグを出現順に $\{e_1^\alpha, e_2^\alpha, ..., e_{N_\alpha}^\alpha\}$，$\{e_1^\beta, e_2^\beta, ..., e_{N_\beta}^\beta\}(N_\alpha < N_\beta)$ とし，各セグメント内のタグ e_i^α と $e_i^\beta (i = 1, 2, .., N_\alpha)$ に対し以下の判定を行う．

- タグ出現パターン
$$x_{\alpha\beta} \leftarrow \begin{cases} x_{\alpha\beta} & (e_i^\alpha = e_i^\beta) \\ Tag\_Correspond\_Value * x_{\alpha\beta} & (e_i^\alpha \neq e_i^\beta) \end{cases}$$
- スタイル (テキスト表現にフォント設定・Bold・Italic 等の指定がある場合)
$$x_{\alpha\beta} \leftarrow \begin{cases} x_{\alpha\beta} & (e_i^\alpha = e_i^\beta) \\ Style\_Correspond\_Value * x_{\alpha\beta} & (e_i^\alpha \neq e_i^\beta) \end{cases}$$
- 色指定 (テキスト表現に色指定がある場合)
$$x_{\alpha\beta} \leftarrow \begin{cases} x_{\alpha\beta} & (e_i^\alpha = e_i^\beta) \\ Color\_Correspond\_Value * x_{\alpha\beta} & (e_i^\alpha \neq e_i^\beta) \end{cases}$$

最終的に得られた値 $x_{\alpha\beta}$ を類似度とし，これが所定の閾値を上回る場合に類似セグメントと判定する．本稿の実験では，類似度 $x_{\alpha\beta}$ の初期値は 1.0 としている．また，$*\_Correspond\_Value$ は 0.8，類似判定のための閾値は 0.6 とした．この2つの値は予備実験に基づいて適切と思われる値を設定した．

以上の処理を適用して決定されたセグメントの例を図3に示す．なお，実線は決定されたセグメントを示し，×は類似と判定され結合したセグメントを示している．図

図 1: 分析の対象としたウェブページ

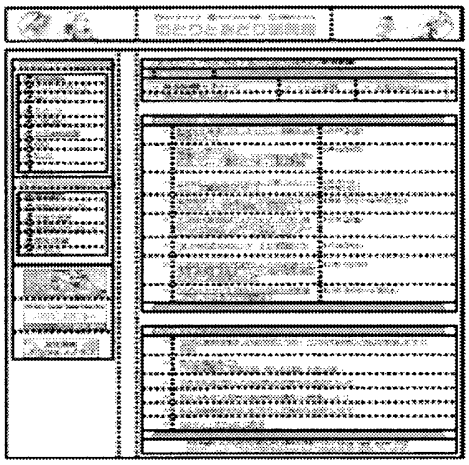

図 2: 初期分割により得られたセグメント

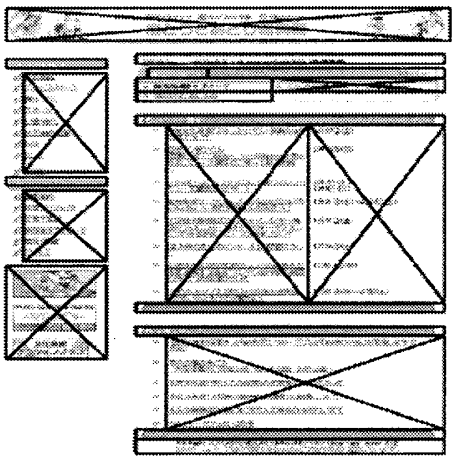

図 3: 最終的に得られたセグメント

表 1: 提案手法によって決定されたセグメントの主観評価 (値は被験者数)

ページ	完全一致	包含	部分包含	不一致
A	**5**	0	0	1
B	0	1	**4**	1
C	2	**4**	0	0
D	0	**6**	0	0
E	**4**	0	0	2

3 からわかるように，タグのみの分割 (図 2 破線で示された小セグメント) では不自然であったセグメント群が，その類似性により 1 つのセグメントに結合され，より自然な形でウェブページが分割されている．

3. 提案手法の評価実験

本提案手法によって決定されたセグメントの妥当性を確認するために，主観評価実験を行った．6 名の被験者に A〜E のページを提示し，アンカーテキストによる記述が最も集中しているセグメントをペンで示してもらい，提案手法より得られたセグメントとの一致の度合いを評価した．その結果を表 1 に示す．

本実験の結果，提案手法により決定されたセグメントと被験者により判定されたセグメントが完全に一致または包含される (自動的に決定されたセグメントが被験者により判定されたそれに含まれる) 割合は約 73 % であった．ページ B については提案手法により 2 つのセグメントと決定された箇所を 6 人中 4 人の被験者が 1 つのセグメントと判断したため，部分包含と判定された割合が高くなっている．以上より，本手法の適用で，概ね人間の感覚に近いセグメントが得られることが確認された．

4. まとめ

ウェブページの有用性を「ページの見え方」に基づいて評価するために，本稿では，ページ上の記述の「まとまり」という特性に着目し，こうした「まとまり」を時動的に決定する手法を提案した．そして，主観的評価実験により，人間による判断と類似した「まとまり」が導出できることを確認した．

本手法により導出される「まとまり」の特性の定量化などを行うことにより，ウェブページの有用性を「ページの見え方」から評価する際の重要な手がかりが得られる可能性がある．今後，本稿での成果を基に，本手法を活用した評価法やウェブページの有用性評価手法に関する検討を進める予定である．

参考文献

[1] Google, http://www.google.co.jp/intl/ja/why_use.html

[2] メッツガー 視覚の法則, 盛永四朗訳, 岩波書店, 1968

[3] 行場 他著, 知性と感性の心理, 福村出版, 2000

[4] ユーザビリティエンジニアリング原論, Jacob Nielsen 著, 三好かおる訳, 東京電機大学出版, 2002

[5] Yodung Yang, et al," HTML Page Analysis Based on Visual Cues", 6th ICDAR2001, 2001