1U-05 HTML文書中の属性記述を基にしたWebページの対比方法

中島 誠　大城 英裕　行天 啓二
大分大学工学部 知能情報システム工学科

1 はじめに

近年のWWW技術の向上により，一般のユーザが各自のもつ情報を発信する機会が増大している．情報を効果的に伝達するには，他者の興味を喚起するデザインで，理解のし易いWebページの作成が望まれるが，経験の浅いユーザにとってこの作業は困難なものとなる．既存の優れたWebページを参考に，ページ作成のための示唆が得られる仕組みがあればその助けとなりユーザのスキルの向上にも役立つ．

本研究の目的は，既存のWebページの中から，ユーザの発想を喚起するものを取り出し，それを基にページ作成を支援する環境の構築にある．本稿では，ユーザに提示する情報を取り出すための方策として，異なるWebページを記述するHTML文書から，記述上の類似点・相違点を明らかにする対比方法について検討する．

2 考え方

Webページ作成の支援では，単にユーザが参考にしたい既存のWebページを提示するだけでは，より斬新な結果は得られない．これまで，複数の構成要素をもつ対象について，別々な対象の要素同士を対応させ，類似点や相違点を明らかにすることが様々な問題解決の場面で役立つとされてきた[4]．ここでは，対象とするWebページ間の局所的にみた構成要素間の記述上の違いと，全体の構成上の差異との両方を明らかにして作成支援のための情報とする．

いま，2つのWebページを取り上げたとして，これらの間の対比を，記述上関連のある構成要素同士をお互いの構造中での他の構成要素との関係を考慮しながら矛盾のないように対応づけ比較することと捉える．ここで対応づけられなかった構成要素は，ページ全体の構成上での相違を示唆し，対応づけられた構成要素同士の記述上の違いは，局所的なWebページの表現方法の違いを示すものとなる．

3 Webページの対比

Webページのような様々な構成の文書からレイアウトの解析によって構成要素を抽出するには多くの場合困難を伴う[3]．ここでは，タグで囲まれたHTML文書中の要素をWebページの構成要素とし，要素に付加されたプレゼンテーションに関連した属性とその値を用いて各構成要素を記述する．

構成要素の記述に用いる属性のうち，色などの記号属性については，その値の間に概念的な関連があるとするのが自然である．また，文字や領域のサイズなど，値が数直線上に並べられる数値属性についても，図1に示すように並んだ値が上位の値としての区間を段階的に形作っていると見られる(例えば1/3は区間[1,3]を意味する)．それゆえ，これらの属性は値の関連が階層構造をしたものと捉えられる．構造中では，各節点は属性値に対応しており，その子孫（先祖）に対しより一般的な（特定的な）関係にある．階層構造を通じて見たとき，より関連の深い値同士を一般化した結果はより特定的になる[1]，このことに注目すると，構成要素間の記述上の関連性は，これらの記述にある属性ごとの値を一般化結果について特定さ具合をみることでわかる．

2つのWebページをSとTとし，対比方法の

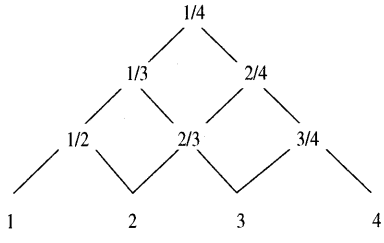

図 1: 数値属性の階層構造表現

Comparing Web Pages by Considering Presentation Attributes of HTML Documents.
Makoto NAKASHIMA, Hidehiro OHKI, and Keiji GYOHTEN
Oita University, 700 Dannoharu Oita 870-1192, Japan

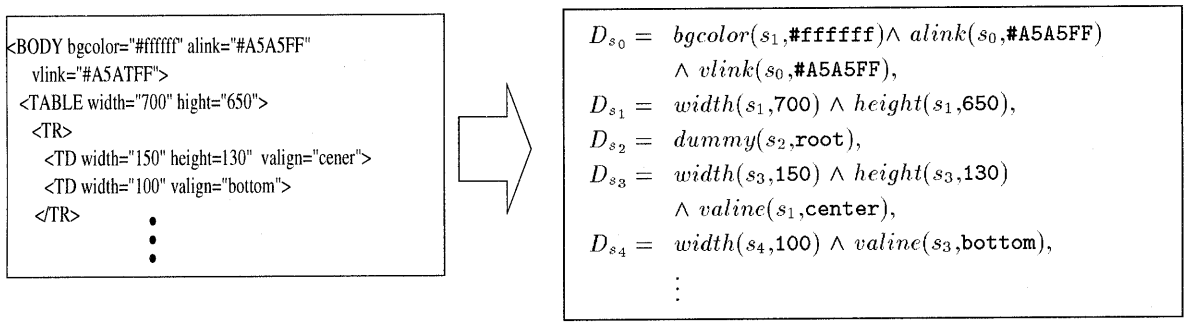

図 2: HTML ソースコードと構成要素の記述例

流れを示す.

(step1) S および T についての HTML 文書から構成要素を抽出しその記述を求める.

(step2) 構成要素同士の対応づけの中から, お互いの構造で矛盾を生じさせない対応づけを全て求める.

(step3) step2 の対応づけの中で, より関連の深い構成要素同士が対応づけられているものを出力する.

4 構成要素の記述の扱い

HTML 文書から抽出した各構成要素の記述として次のような属性 $A,\ldots,B$ についての 1 項述語の連言を用いる.

$$D_{s_i} = A(s_i,\delta_i^A) \wedge \cdots B(s_i,\delta_i^B)\,(\mathrm{i}\geq 1)$$

ここで, δ_i^A は構成要素 s_i がもつ属性 A の値である. 各構成要素の記述の組によってページ全体が表される. HTML 文書中の要素で属性とその値が欠損している場合には, その要素が含まれる要素に付加された属性とその値を利用できる範囲で (例えば BODY 要素の属性 bgcolor は BODY 要素に含まれる要素についても反映される) その記述を補完する. それ以外の場合には, 便宜的な仮の属性と値を使って記述しておく. 図 2 に, HTML 文書とそれから得た構成要素の記述の例を示す.

5 対比方法の効率化

2. の対比の方法で矛盾を生じない可能な構成要素同士の対応づけを全て求めることは, 対象とする Web ページが大きくなると現実的でない. これには, 構成要素同士の対応づけを節点とする連合グラフを利用して [3], 最も多くの構成要素同士が対応させられるような対応づけを選ぶようにする. 上記連合グラフ上では, 互いに矛盾しない節点の最大の組み合わせ (複数の場合もある) がクリークを形成することから, 数学的に求められる.

また, 上で得られた対応づけが複数の場合, 構成要素同士の記述を一般化した結果の特定さ具合を数量化したレベル総和 [2] が大きくなる対応づけを選んで残すようにする. レベル総和は, 属性の値を節点とした階層構造中で, 下にある節点ほど高いレベルを設定し, 各対応づけで, 対応づけられた構成要素の記述の一般化結果に現れる値のレベルを合計したものである. それゆえ, レベル総和の大きくなる対応づけのうちから, 関連深い構成要素同士が対応させられているものを見つけるようにする.

6 おわりに

HTML 文書に書かれた情報を基に, Web ページを構成要素に分割し, 異なるページ間での記述の相違点を見つける方策について提案した. 今後は, システムへの実装と, 見つけた相違点を効果的にユーザに提供する視覚化法の実現が課題として残されている.

参考文献

[1] Dietterich, T. G., and Michalski, R. S.: A comparative review of selected methods for learning from examples, R.S. Michalski et al. (eds.), Machine Learning, Morgan Kaufmann, 1983.

[2] Kolodner, J. L.: Case-Based Reasoning, Morgan Kaufmann, San Mateo, CA, 1993.

[3] 石谷 康人: 創発的計算に基づく文書画像レイアウト解析, 画像の認識・理解シンポジウム, MRIU96, pp.343-348, 1996.

[4] 中島 誠, 伊藤 哲郎: 背景知識を解した概念構造の対比操作, 人工知能学会, Vol.13, No.6, pp. 990-1001, 1998.