1C4-01

視覚的類似性に基づくWebページ検索手法の提案

Finding Web Page Based on Web Page Similarity as Image

三橋　憲晃　　山口　亨　　高間　康史
Noriaki Mitsuhashi　Toru Yamaguchi　Yasufumi Takama

東京都立科学技術大学
Tokyo Metropolitan Institute of Technology

When we search information on the Web using search engines, they only analyze the text information collected from the source files of Web pages. However, there is a limit to analyze the layout of a Web page only from its source file, although Web page design is the most important factor for a user to estimate a page. In particular, it often happens on the Web that the pages of similar design offer similar information. We propose a method to compare the design of pages by treating the displayed page as image.

1.　はじめに

現在，インターネットの普及により誰でも簡単にネットワークを使って情報収集を行うことができる．ネットワーク上の情報は膨大であり, 必要としている情報を検索する手法が重要である．ネットワーク上で情報検索を行う際, Google などの検索エンジンでは, Web ページのソースファイルから抽出したテキスト情報を用いている．しかし, ソースファイルからテキスト情報を集めるだけでは, Web ページのレイアウトや表示された時の様子を解析するには限界がある．実際には Web ページを見る際, 人間はソースファイルの情報ではなく画面上に表示されたものを視覚的に捉え情報を集めていると考える．そこで本稿では，類似した情報を提供している Web ページでは, デザインが似ている場合が多いことに注目し, Web ページを視覚的に画像として扱う事によりデザインの比較を行う方法を提案する．提案手法により，トップページなどの特徴的なページを判定可能である事を実験により示す．

2.　web ページ比較方法の概要

Web ページ比較方法の流れを図 1 に示す．まず, 目的の Web ページを画像に変換し, エッジを検出するために輝度成分である Y 画像と色差成分である Cb,Cr 画像へ変換を行う．輝度成分は, 人間の視覚にあたえる影響が大きいため Y 画像を用いて処理を行う．Y 画像を n×n 画素ブロックに分割し, ブロックの平均値を求め，ブロック平均値で各画素値を割った値を用いてエッジ検出を行い，2 値化された画像を抽出する．2 値化された画像をもとに文字・画像などの分類を行う．

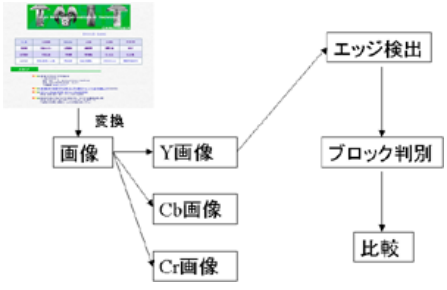

図 1: ページ比較方法の流れ

連絡先: 三橋　憲晃, 東京都立科学技術大学, 東京都日野市旭が丘６－６, E-mail:noriaki@krectmt3.tmit.ac.jp

3.　レイアウト解析方法

Web ページを比較する前処理のレイアウト解析法について説明する．本節では，文字・画像を一つのブロックとして分割し，判別する方法を述べる．

3.1　エッジ検出

まず，エッジ検出を行う前に対象画像を RGB 空間から YCbCr 空間へと変換を行い Y 空間を用いて処理を行う．これはエッジ検出を行う際，RGB すべての値に対し処理を施すより Y 空間だけで処理を行ったほうが効率的だからである．エッジ検出では，Y 画像を 2×2 画素ブロックに分割し，それぞれのブロックで平均値 X を算出する．算出されたブロック平均値 X で各画素値 x_i を割った値 α_i を求める．

$$\alpha_i = \frac{x_i}{X} \quad \begin{array}{l} x_i : \text{画素値} \\ X : \text{ブロック平均値} \end{array} \tag{1}$$

この α_i の平均値は，隣接画素間の相関性の高さから $\alpha_i = 1.0$ に集中するため，α_i はブロック平均値に対する急峻性を表していることになる．つまり，$\alpha_i = 1.0$ が平坦部であり，$\alpha_i = 1.0$ から離れると急峻性が高くなる [1][2]．その特徴を利用し，α_i がある閾値を超える場合エッジ部と判定しエッジ検出を行う．

3.2　Web ページにおける領域分割法

エッジ検出を行い２値化された画像を用いて，文字・画像を領域分割し一つのブロックとして抽出する．Web ページでは文字・画像などが混在しているが，文章や画像など一つのまとまった領域は，文字・画像のエッジ間の距離が小さい．逆に，文章と画像など違った情報の領域は同じ種類の領域に比べエッジ間の距離が大きい．今回は，エッジ間の距離 d を用いて領域分割を行った．領域分割を行う際に閾値 a を定め，距離 d が閾値 a を超えるかどうかで同じ領域に属するかの判定を行う．まず，画像をラスター方向に検索していき最初に検出したエッジ部の位置情報を記録する．その位置情報を基に，上下左右エッジを検索していき $d \geq a$ となるまで処理を行う．この処理を行いながら位置情報の最大値と最小値を記録し，再帰的に繰り返すことにより一つの領域をブロックとして検出する．また，背景色の変化などによるエッジや閾値の値によって目的の領域より大きいサイズで領域分割されてしまうことがある．この問題を解決するために，一定の大きさを超えるブロックでは，さらにそのブロック内で閾値 a の値を小さくし領域分割行う．この操作を施すことにより正確に領域を検出しやすくする．

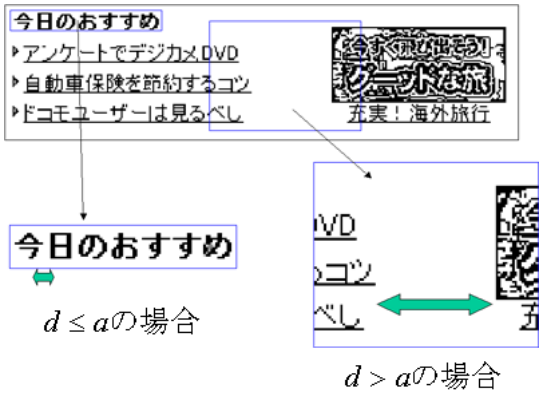

図 2: 領域分割法

3.3 文字・画像の抽出

3. 2 節において，領域分割によって抽出されたブロックは，その内容が文字・画像のいずれかであるかは不明であるが，Web ページを比較する上で，文字・画像（自然画像・イラスト）の判別を行うことは重要である．本節では，Web ページ上の文章は基本的に横書きということを仮定して文字・画像の判別を行う．まず，それぞれのブロックに対しブロック内に文字が含まれているかどうかの判断を行う．文字列を含むブロックは図 3 に示すように，ラスター方向に検索をしていくとエッジが検出されるピクセル列と検出されないピクセル列が現れる．文字列を抽出する際，最初にエッジが検出されたピクセル列の位置情報を記憶しておく．そして，最初に検出されたピクセル列からエッジが検出されないピクセル列までの位置情報を記録しておき，その領域で文字かどうかの判断を行う．図 4 に示すように，一定の間隔でエッジが連続して出現した場合文字と判断する．また，画像やイラストの場合ブロック内では不規則にエッジが出現したり，エッジが検出されないピクセル列がなかったりする．その場合，そのブロックは画像・イラストと判断する．この処理を行うことで，抽出されたブロックの属性を決めることができる．

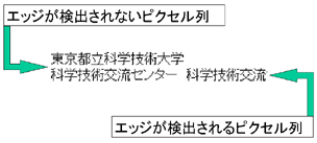

図 3: 文字列検出

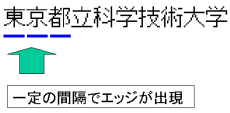

図 4: 文字判断

4. Web ページの比較方法

抽出されたブロックのレイアウトによって Web ページを比較する方法を提案する．Web ページから抽出されたブロックの属性は文字・画像・文字と画像の 3 つに定め，それぞれ重み付けを行い Web ページを 1 つのパターンとみなす．重み付けは正の値を用い，文字を 1，画像を 2，文字と画像を 3，ブロック以外を 0 と設定し計算を行った．比較する 2 つの Web ページのパターンを，$\mathbf{w_i}$, $\mathbf{w_j}$ とベクトルで表現した時の距離 S_w は，式 2 で表される．ここで，$\mathbf{w_i} \cdot \mathbf{w_j}$ は内積であり，$\| \mathbf{w_i} \|$ と $\| \mathbf{w_j} \|$ は大きさを示している．

$$S_w = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\| \mathbf{w_i} \| \cdot \| \mathbf{w_j} \|} \qquad (2)$$

S_w は単純類似度と呼ばれる．その値は 0 ～ 1 をとり，1 に近づくほど比較した 2 つの距離が近いことを示している [3][4].

5. 結果

図 5 に Web ページの原画像を，図 6 に Web ページを領域分割した結果を示す．また，Web ページの比較については，表 1 と表 2 に示す．領域分割では，図 6 からわかるようにおおむね正確にブロック分割を行うことができている．表 1 と表 2 では，検索サイトである goo(http://www.goo.co.jp) と他の Web ページとの比較を示しているが，レイアウトによって単純類似度を算出し比較を行った結果と視覚的な判断を行った結果において，ほぼ同等の結果を得ていることがわかる．

表 1: 単純類似度

ページ名	類似度
yahoo!	0.522
infoseek	0.503
onkyo	0.300
daihatsu	0.299

表 2: 視覚的評価

ページ名	視覚的判断
yahoo!	6/7[人]
infoseek	7/7
onkyo	3/7
daihatsu	1/7

図 5: 原画像

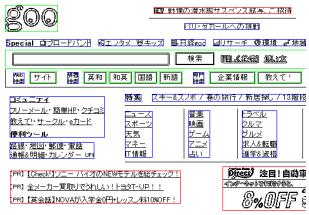

図 6: 領域分割画像

6. まとめ

6.1 結論

本稿では，Web ページを視覚的に画像として扱う事によりデザインの比較を行う方法を提案した．Web ページを画像として扱い，領域分割を行って目的のブロックを抽出することができた．さらに，レイアウトを解析し類似度を計算することによって Web ページの比較を行うことができた．また，Web ページ比較の結果と人間の評価ついてほぼ同一の結果を示せた．今後の課題として，カラーヒストグラムなどを用いた文字・画像抽出の精度の向上や Web ページの比較に背景色などを用いる事が考えられる．また，レイアウトの比較方法としては，ブロックの位置・サイズを考慮した相対的な比較方法で，さらに精度を上げることが期待できる．

参考文献

[1] 三橋憲晃，佐藤和弘　「パターン係数ブロック判別法によるベクトル量子化画像圧縮の検討」2001 年　東京都立科学技術大学学士論文

[2] 松坂　健治，佐藤和弘　「凹凸係数を用いた画像処理アルゴリズムの検討」2001 年　東京都立科学技術大学修士論文

[3] 谷口慶治　「画像処理工学　-基礎偏-」　共立出版 (1999)

[4] 谷口慶治　「画像処理工学　-応用偏-」　共立出版 (1999)