2008－DBS－146 (53)
2008／9／23

色情報を利用したWebページ検索手法について

能美　礼†　大野　成義††　石川 博†††
†職業能力開発総合大学校　電気・情報専攻　〒229-1196　神奈川県相模原市橋本台4-1-1
††職業能力開発総合大学校　〒229-1196　神奈川県相模原市橋本台4-1-1
†††静岡大学　情報学部情報化学科　〒432-8011　静岡県浜松市中区城北3-5-1
E-Mail: †m19513@uitec.ac.jp, ††ohno@uitec.ac.jp

あらまし 本研究では過去に閲覧したことのあるWebページや, 人から伝聞したWebページなどページ自体のイメージが検索者側にある場合に使用する検索システムを提案する．色情報とその配置位置により Webページを検索する方法であるため，一般的なイメージ検索と異なりこのシステムではキーワードを必要としない．検索対象となるのはWebページそのものではなく各配置位置における索引色の割合というデータである．提案したシステムの有効性を確認するためにプロトタイプを実装し実験を行った．

キーワード 情報検索　画像処理

A Method of the Web Page Retrieval Passed on the Color Impression

Rei NOUMI†　Shigeyosi　OHNO††　Hiroshi　ISHIKAWA†††
†Polytechnic University　Hashimotodai 4-1-1, Sagamihara-shi, 229-1137 Japan
††Polytechnic University　Hashimotodai 4-1-1, Sagamihara-shi, 229-1137 Japan
†††Faculty of Informatics, Shizuoka University Jouhoku3-5-1, Hamamatsu-shi, 423-8011 Japan
E-Mail: †m19513@uitec.ac.jp, ††ohno@uitec.ac.jp

Abstract In this paper, we propose a search engine used when the searcher have the image of Web page, or perused the Web page in the past. Since it is how to search a Web page using color information and its arrangement position, unlike general image search, a keyword is not needed in this system Our search system uses the rate of the index color in each area instead of the Web page itself. This paper also shows experiments in order to confirm the validity of the proposed system.

Key words Information retrieval, Image data processing

1. はじめに

今までインターネット上で必要な情報を探すために,Webページが文書データであったことからキーワード検索が主に用いられてきた．

現在ではWebページはテキストだけでなく画像などコンテンツが多彩になりマルチメディアデータ化してきているがこれらの検索もキーワードを用いられている．しかし常に的確なキーワードが思い浮かぶとは限らない．検索したいものを色や形体といったイメージで検索したいという状況が存在する．

過去に見たことのあるWebページや伝聞したWebページなどページ自体のイメージが検索者側にある場合それを用いた検索方法を提案する．この検索方法はキーワードを必要としない．

視覚的印象を使用した検索の例として限られた絵画やイラスト，写真などを対象とした画像検索が存在する．これをWebページ検索に応用する．

従来の画像検索の方法は画像をDBに保存し, それを元に検索を行うというものである. しかしWeb

ページの検索に用いる場合，その数は膨大であり，更新も頻繁に起こるため画像を DB に保存することに無理が生じる．そこで画像自体を保存するのではなく画像の色がどのような割合で含まれているのかという情報とその配置位置のデータのみを保存することで DB の軽量化と検索速度の向上ができるのではないかと考えた．本研究では色情報とその配置位置による Web ページ検索システムを構築しその高速化を図る．

2. 色とその配置位置によるWeb検索システム

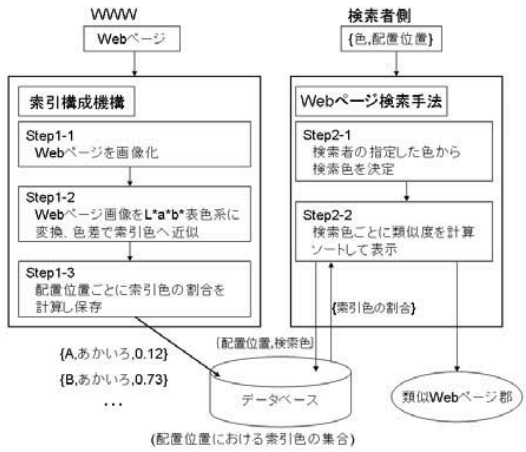

図 1　提案手法の全体構成

色と配置による Web 検索システムを実現させるためには以下のような方法をとる．

Step1．索引構成機構を用いて DB を作成する．(図 1 の左側部分)

Step1-1．検索者が記憶している部分は実際に表示された部分であることから，ブラウザではじめに表示された部分のみを画像として保存する。また最近ではフラッシュを使用しているサイトが多いことを考慮して完全に表示されてから 5 秒後のブラウザに表示された画像を保存している．

Step1-2．保存した Web ページ画像の 1 ピクセル毎の RGB を L*a*b*表色系に変換する. 変換したデータを 1 ピクセルごとに索引となる色へ近似する．この DB 化のための色を索引色と呼ぶ．L*a*b*表色系は RGB からの変換式が用意されており色差が一定の空間になるため，色の違いが距離で表現可能となる利点を持つ．以下の変換式を用いて RGB 値を L*a*b*値へ変換する[3].

$$X = 0.3933R + 0.3651G + 0.1903B \quad (1)$$

$$Y = 0.2123R + 0.7010G + 0.0858B \quad (2)$$

$$Z = 0.0182R + 0.1117g + 0.9570B \quad (3)$$

$$L^* = 116\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16 \quad (4)$$

$$a^* = 500\left[\left(\frac{X}{X_n}\right)^{\frac{1}{3}} - \left(\frac{Y}{Y_n}\right)^{\frac{1}{3}}\right] \quad (5)$$

$$b^* = 200\left[\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - \left(\frac{Z}{Z_n}\right)^{\frac{1}{3}}\right] \quad (6)$$

ここで式(4)，(5)，(6)における X_n, Y_n, Z_n は Web ページ閲覧時の環境により異なる刺激値を補正する定数であり，文献[3]に従い ，

$X_n = 98.072, Y_n = 100, Z_n = 118.225$ とした.

任意の 2 色 $c_1(L_1, a_1, b_1), c_2(L_2, a_2, b_2)$ の色差 $\Delta\dot{E}(c_1, c_2)$ は式(7)で計算可能である.

$$\Delta\dot{E}(c_1, c_2) = \sqrt{(L_1 - L_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2} \quad (7)$$

式(1)-(6)を用いて，Web ページ画像に含まれる全画素を RGB 値に変換する. その後、式(7)によって索引色との色差を計算して近似する. 提案手法では配置位置に含まれる索引色に対して検索者が選択した色との色差を用いて重み付けを行う. ある索引色 c_1 が選択されたとき c_2 につける重み $\Delta\dot{E}(c_1, c_2)$ を次のように定義する.

$$\Delta E(c_1, c_2) = \frac{\Delta\dot{E}}{(\text{索引色間距離の最大値})} \quad (8)$$

Step1-3．Web ページの特徴として左部にメニューボタンを並べたり，上部に企業の色やサイトのタイトルなどをいれるなどある程度規則性のあるレイアウトを持つ．このことから画像を全体の 25%の面積を持つように上下左右中心の 5 つに分割する．それぞれ図 2 のように ABCDE とする．分割された 5 つの各配置位置に含まれる索引色の割合をデータベースに保存する．

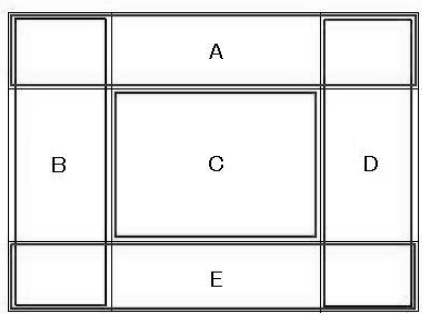

図 2　レイアウトの特徴からなる配置位置

Step2．検索するときに検索者は色と配置位置を決定する．(図 2 の右側部分)検索者の指定した色を検索色と呼ぶ．

Step2-1．各配置位置の検索色を指定する．この時選択できる検索色を{赤，緑，青，オレンジ，黄，茶，桃，白，黒}の 9 色としている．

Step2-2．検索色と索引色の色差を利用して検索者が入力した検索色と配置位置から Web ページの類似度 *Sim* を求める．

Sim は各配置位置に含まれる各索引色の割合を元に構成される α 類似度と，検索者が指定していない配置位置における各配置位置に含まれる各索引色の割合を元に構成される β 類似度を用いて次の式で定義する．

$$Sim = \alpha + g \cdot \alpha \cdot \beta \qquad (1)$$

α類似度は検索者の記憶している部分に対応しβ類似度は記憶していない部分に対応する．また g は重み付け係数である．g が 0 のとき，検索色を指定しなかった配置位置部分を無視して類似度を決定する．g が大きくなれば何も指定していない配置位置には他の位置で指定した検索色は含まれないはずだという過程で類似度が決まる．上記 2 点に基づき検索者が k 番目の検索色を選択した場合，i 番目の配置位置における α 類似度 αik, βik は次の式で定義する．

$$\alpha_{ik} = \sum_{j=0}^{allC-1} C_{ij} \cdot (1 - \Delta E(CL_k, CL_j)) \qquad (2)$$

$$\alpha'_{ik} = \sum_{j=0}^{allC-1} C_{ij} \cdot (1 - \Delta E(Cl_k, CL_j))^2 \qquad (3)$$

$$\beta_{ik} = \sum_{j=0}^{allC-1} C_{ij} \cdot \Delta E(CL_k, CL_j) \qquad (4)$$

allC，*CL k* はそれぞれ索引色数，*k* 番目の索引色を表す．(3)式は類似度に色差を強く反映させるため重みとして色差を二乗している．また，検索色を 9 色に限定したため幅を持たせる．この幅を(色差)<0.2 のとき *ΔE* の値を 0 とした．

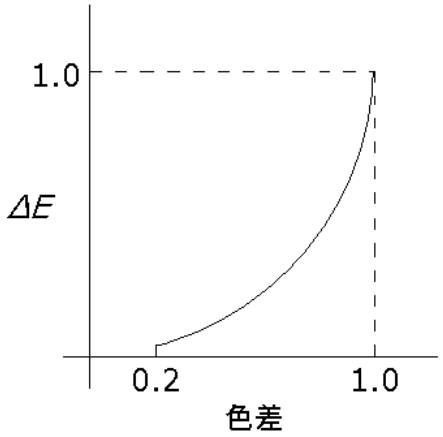

図 3　誤差の範囲を考慮した色差

式(3)(4)を Web ページ全体に適用した α，β は次の式で定義する．

$$\alpha = \sum_{i=0}^{allP-1} \sum_{k=0}^{allC-1} \alpha_{ik} \cdot ChC_k \cdot ChP_i \qquad (5)$$

$$\beta = \sum_{i=0}^{allP-1} \sum_{k=0}^{allC-1} \beta_{ik} \cdot ChC_k \cdot (1 - ChP_i) \qquad (6)$$

allP は Web ページの分割数である．また，*ChCk*, *ChPi* は検索者が指定した検索色，配置位置であるかどうかを表し，選択していれば 1，そうでなければ 0 とする．

式(1)を索引構成機構によって生成された全索引に対して適用し，類似度 *Sim* の大きい順にソートして表示する．

3. 検索の高速化

DB に格納された Web ページのデータが多くなると類似度を計算してソートをするのに長い時間が必要となる. 検索要求があってから検索結果を返すまでのレスポンス時間を短縮するために, 図 1 の Step2-2 の計算を事前に行うように修正する. つまり Step1-3 の直後に想定される検索色すべてに対する類似度の計算を行う.

4. 実験, 考察

提案方法の有効性を確認するために実験を行った. 索引色は JIS 規格で定められた色鉛筆の色をベースに 43 色[4]で構成した. 指標となる Web ページ画像は無作為に選んだ 200 ページ分の画像を用いた.

実験 1 配置位置全箇所を同一色に指定した実験

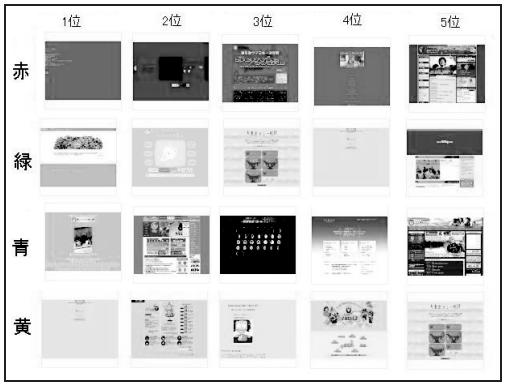

図 4 実験 1 全箇所を同一色で求めた結果の例

5 つの配置位置のすべてを同じ色を指定して検索を行った. その実験結果が図4である.

青, 黄に関しては申し分のない結果である. 赤の検索結果の上位に青いページがまぎれているのはサムネイル上では青に見えるが実際には色が赤と色差の小さい赤紫であるためにこのように上位に検索された結果となっている. 緑の検索結果には緑と黄の色差が小さいために黄が混ざりこんでいる. またこのことから, 黄と白の色差が小さいことため黄を選択した際検索結果に白の多いページが含まれることを心配した. しかし, 白の多いページが検索結果の上位に現れることはなかった.

実験 2 αik と $\alpha' ik$ の比較実験

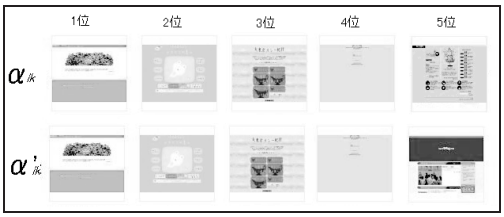

図 5 実験 2 αik と $\alpha' ik$ の比較実験

(3)式で与えた重みによって検索結果にどのように影響するか(2)式との比較実験を行った. 配置位置を上部のみに緑を指定して検索した場合, 図 4 のように $\alpha' ik$ の方が強く特徴が現れている. 定量的に評価するために DB の Web ページ画像数をもっと多くする必要がある.

実験 3 重み付け係数 g の妥当性についての実験

図 6 検索色を青に指定

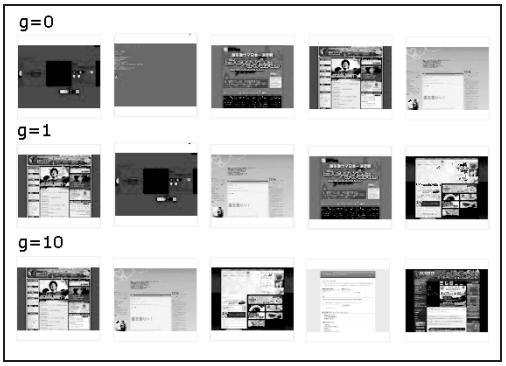

図 7 検索色を赤に指定

検索色を赤, 青とし配置位置を上部のみとした場合で検証した. その結果が図 6, 図 7 である.

g=0 のとき検索色を指定していない配置位置にも検索色の含まれるページが上位に現れている. これと比べて g=1 のときの検索結果である Web ペ

ージの配置位置以外の箇所に指定した検索色とは違う色が特徴として現れていることから g は妥当であると言える．また g の強弱に関して g=10 の方が g=1 よりも特徴が強く現れている．

実験 4 配置位置による変化

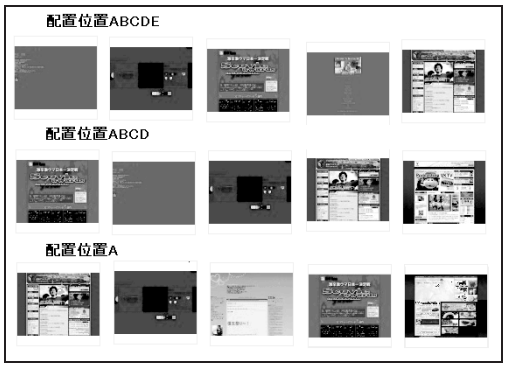

図 8　検索色を赤に指定

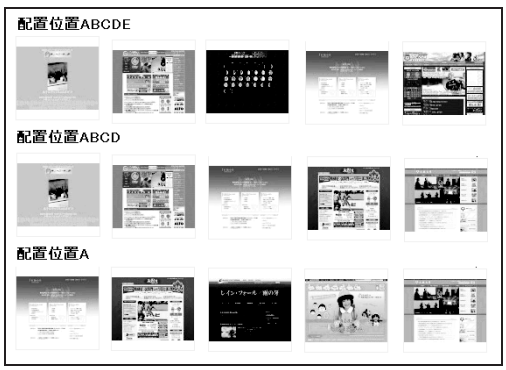

図 9　検索色を青に指定

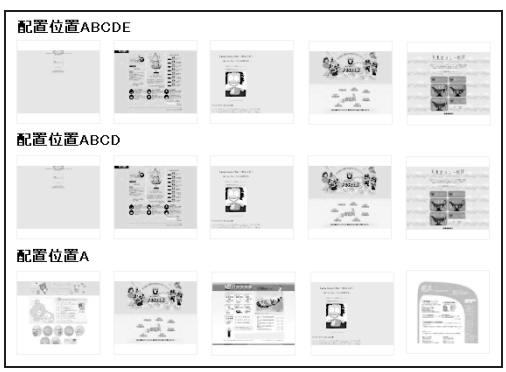

図 10　検索色を黄に指定

図 11　検索色を緑に指定

検索色の配置位置を減らしていく(全体→下部以外→上部のみ)．この時，慣性係数 g= 1 としている．その実験結果が図 8-11 である．

全体を配置位置とした場合と配置位置を下部以外とした場合は結果が似たものになっている．配置位置を上部のみにした場合では全体を検索色としたときとは結果が明らかに異なるものになった．

実験 1 の説明でも述べたとおり検索色を緑に指定した場合全体的に黄色いページが上位にランクされてしまう印象がある．

配置位置の変化による検索結果として配置位置の変化に対応しているためシステムとして期待したようにと動作していることがわかった．

5. まとめと今後の課題

色差の小さい緑と黄のように結果が混ざり合う場合もあるが，ほぼ期待通りの結果が得られるシステムの構築が出来た．

色差への重み付けは効果がみられた．重み付け係数 g の強弱によって検索結果に選択した検索色の強調に強弱がつけられる．

今後定量的評価を行うためにサンプルデータを最低でも数倍にする必要がある．

検索結果の表示は 1 位から順に表示するのではなく，検索時の指定配置位置を考慮した 2 次元的表示システムを検討している．検索者にとって見やすい表示である必要があるため被験者を募って今後見やすさや使い易さを確認する実験を行う．

6. 参考文献

[1]石川幹直　細川宜秀　高橋直久，色とその配置位置に基づいた視覚的印象による Web ページ検索手法の実現方式，　DEWS2005

[2] Erwan LOISANT,Hiroshi ISHIKAWA, José MARTINEZ,Manabu OHTA,Kaoru KATAYAMA, User-Adaptive Navigation Structures for Image Retrieval ,DBSJLetter

[3]名阪カラーワーク研究会 http://www005.upp.so-net.ne.jp/fumoto/index.htm

[4]めざら資源 http://homepage3.nifty.com/mezala/pc/jiscolor/jiscolor.html

[5] 小屋夕介 中西崇文 北川高嗣，L*a*b*表色系を利用した静止画像からのメタデータ自動抽出方式，DEWS2003