L-082

キーワードと画像特徴を利用した Web ページ検索システム

A Web Page Retrieval System Based on Keywords and Image Features

松久保　美幸 †　　深海　悟 †
Miyuki Matsukubo　Satoru Fukami

1. まえがき

インターネット上における爆発的なコンテンツの増加に伴い，検索エンジンに対する要求も高まっている．特に Web ページを大量に収集しインデックスを作成して検索サービスを提供する形式のサービスが増加している．

しかし現在一般的である検索エンジンではキーワードを基づくものが主流である．そのため過去に閲覧したページのアドレスがわからず，かつ的確なキーワードが連想できない場合など，ユーザの意図する有効な検索結果が得られないことが多い．

例えば「以前に見たフリーソフトがたくさん紹介されていたあの緑色のサイトが見たい」などの場合，「フリーソフト」というキーワードを基にユーザが所望するサイトを検索することは困難である．

問題の解決手段としてはいくつか考えられるが，本稿では，Web ページから画像特徴量を抽出し，ページ間の距離を算出することにより，視覚的な情報を考慮に入れた Web ページ検索システムを提案する．

2. 目標とするシステム

本システムで目指すシステムは以下の通りである．

- WWW 上にある Web ページを対象に視覚による検索が行えること
- ユーザがキーワードの入力とマウスによる操作のみで検索が行えること（ブラウジングによる検索）
- 高速な検索が行えること

本システムではユーザが気軽に検索できるようにするため，例示画像を提示することなく，平易なキーワードと簡単な操作だけで高速に検索が行えることを目指す．

3. システムの概要

本システムは大別して以下の 3 つに分けることができる（図 1）．

- サムネイル収集部
- Web ページ分類部
- 可視化部

Web ページの検索を行うためにはあらかじめ大量のデータを収集し，インデックスを作成する必要がある．そこでサムネイル収集部では Google [1] に代表される商用のテキスト検索エンジンを利用し，結果として得た URL より上位から順に HTML ファイルを取得し Web ページのスクリーンショットを収集する．

Web ページ分類部では収集した画像データから画像特徴量を計算し，類似度を算出する．

可視化部では算出した類似度を基にユーザが直感的に選択できるよう可視化を行い，ブラウジングによる提示を行う．

ユーザは，はじめに入力フォームにキーワードを入力する．そして得られた検索結果の中からユーザのイメージしているページと似ているページを順次選択していく．ユーザがサムネイルをダブルクリックすると　対象のページが Web ブラウザ上に表示される．

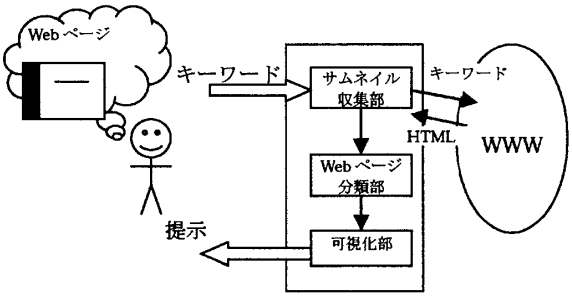

図 1 本システムの概要

4. システムの実装

4.1 サムネイルの取得方法

本システムでは Web ページのサムネイルを取得する手段としてフリーソフトとして公開されている url2bmp[2] を採用した．url2bmp は指定した URL の Web ページのスクリーンショットを保存することができる．

4.2 画像特徴量の抽出

画像の色分布を表す特徴量として Rubner らによって提案された Color Signature を用いる．一般的な Color Histogram ではビン (bin) のサイズが固定であるため効率性と表現力との兼ね合いが難しい点と，histogram 間で定義される距離が人間の感覚と合わないという問題点がある．Color Signature では対象に応じて要素数をかえることができより柔軟な表現が可能になる．

取得した画像は 120x90 の大きさに正規化し，縦と横をそれぞれ 3 分割し合計 9 個の格子状のブロックに分割する．各ブロックの画像データの色情報を一旦 CIEXYZ 色空間に

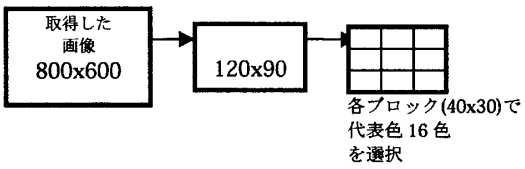

図 2 カラーシグネチャの抽出

† 大阪工業大学
Osaka Institute of Technology

変換したのち，均等色空間(UCS)の一種である CIE1976(L*a*b*)色空間に変換し，各軸につき 6 分割し 216 色に減色を行った．その後各ブロックから要素の割合が多いものから順に最大 16 色を選択し，代表色の i 番目の色 p_i，ブロック画像内の要素数の割合を重み w_i として Color Signature を構成する(図 2).

各ブロックの特徴量を P としたときの Color Signature は次式(1)で表される．

$$P=\{(p_0,w_{p0}),..,(p_{m-1},w_{p(m-1)})\} \quad (1)$$

4.3 Web ページ間の類似度の算出

Color Signature 同士の距離の算出には Earth Mover's Distance (EMD)[3] を採用する．EMD は線形計画法のヒッチコック型の輸送問題の解を用いて計算される手法であり，ミンコスキ形式距離や 2 次形式距離よりもより人間の感覚に合った結果が得られる．また Color Signature 同士の要素数が異なる場合でも距離を算出することが可能である．Color Signature P と Color Signature Q 間の距離 $EMD(P,Q)$は次式(2)を用いる．

$$EMD(P,Q)=\frac{WORK^*(P,Q)}{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}f_{ij}} \quad (2)$$

ここで $WORK^*(P,Q)$ は輸送問題の総コストを最小にする最適解である．$WORK^*(P,Q)$ では総量が少ない Color Signature が選ばれる可能性があるため総フローで割り正規化を行う．

4.4 類似度の可視化

Web ページのサムネイルを類似度が反映された 2 次元空間上に配置し，ブラウジングに基づく検索を行う．

見た目が似ているページは近くに配置し，そうでないものは遠くに配置する．また大量の Web ページを一度に可視化するとサムネイル同士が重なり，可視性に問題が生じる．

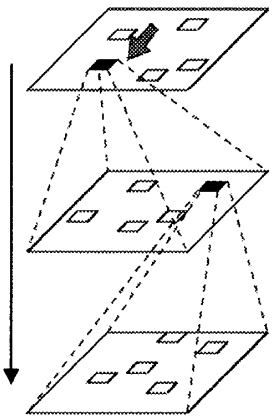

図 3 代表画像の順次選択

そこで検索対象となる Web ページ群の算出した類似度に対し階層的クラスタリングを行い，クラスタの代表ページとなるサムネイルを表示する．ユーザが似ていると思ったサムネイルを順次選択していくことで下位層へと下り，選択したサムネイルのクラスタ内でより精度の高い検索結果が得られるようにした．

高次元空間の可視化には線形手法に基づくものや，ばねモデルに基づくものなどさまざまな次元縮約手法が提案されているが，本研究では視覚度の精度が高い線形手法に基づく古典的 MDS を採用した．本システムでキーワード「apache」を入力し検索した結果を図 4 に示す．

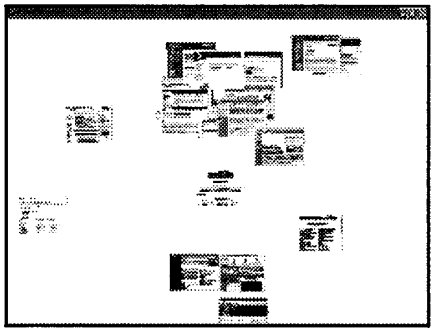

図 4　キーワード「apache」を入力した時の結果

5. 関連する研究

Web ページを画像として扱い類似性を求める研究としては三橋らが提案したレイアウトによる検索手法[4] や橋本らによるオブジェクトの空間的な関係を考慮した検索手法[5] が存在する．どちらもユーザが描いたスケッチ画像を例示画像としている点，色情報に関する尺度が考慮されていない点が本システムと異なる点である．

6. おわりに

本研究で提案するシステムにより，利用者がキーワード入力と簡単な操作だけで，視覚的印象より Web ページの検索を行うことができた．

しかし，サムネイル収集部においては Web ページのレンダリング速度がボトルネックになっており，多くの時間を要するため，これを高速化する手段を考える必要がある．また可視化部においてもクラスタリング手法やユーザインターフェイスの面には改善の余地があり，今後の課題として検討していきたい．

謝辞

本研究にあたり，懇切丁寧な指導を賜りました牧俊男氏に感謝いたします．

参考文献

[1] Google: http://www.google.com
[2] url2bmp:
http://www.pixel-technology.com/freeware/url2bmp/english/
[3]Y. Rubner and C. Tomasi. : Perceptual Metrics for Image Database Navigation. Kluwer Academic Publishers, Boston, December, (2000)
[4]三橋憲晃，山口亨，高間康史：視覚的類似性に基づく Web ページ検索手法の提案．第 17 回人工知能学会全国大会，(2003)
[5] 橋本泰成，五十嵐康夫：レイアウトによる WEB ページ検索．インタラクション 2004，情報処理学会，(2004)