I_066

視覚情報に基づくWebページ閲覧履歴検索

View-based Web Page Retrieval from the Browsing History

渡井 康行 Yasuyuki Watai　中平 浩二 Koji Nakahira　山崎 俊彦 Toshihiko Yamasaki　相澤 清晴 Kiyoharu Aizawa

1. はじめに

Webページは人がWebブラウザを用いて閲覧することを前提に作成されている。人が閲覧する際には、ページに含まれる文章、画像、映像などのコンテンツと同様、Webページそのもののデザインも印象を決定付ける重要な要素といえる。実際に、人は以前に見たことのあるページや、類似するページを一瞥しただけで識別できる。

筆者らは、画像内容検索(Content-Based Image Retrieval: CBIR)をWebページの検索に応用し、Webページを配色、レイアウトなどの見た目の情報に基に検索、キーワード以外の手段でWeb上の情報へのアクセスを可能にする手法の開発に取り組んでいる。このような手法を用いることで、あいまいな記憶を元に以前に訪れたページを再訪することが容易になる。本稿では特にこの再訪シナリオに特化したアプリケーションとして、ユーザのPC上で動作するWeb閲覧履歴検索システムの要素技術について報告する。

2. システムの概要

我々の提案するシステムの概要を図に示す。本システムはユーザのPC上で動作するHTTPプロキシサーバであり、ユーザが閲覧しているWebページをバックグラウンドで解析して特徴量を抽出し、データベースに格納する。

ユーザは簡易なスケッチをクエリとして、システムに蓄積された履歴から所望のWebページを検索することができる。本システムでは入力によって変化の生じた部分だけを用いて履歴内のページとクエリとの類似度を算出し、逐次的に絞込みを行う。

視覚的特徴はWebページのスクリーンショットから抽出する。マッチングのための特徴量としてはカラーシグネチャ[1]を採用する。カラーシグネチャは、色ベクトルとその重み(例えば、画像に占める面積)との組で表される特徴量である。カラーシグネチャ同士の類似度は、Earth Mover's Distance (EMD)を用いて計算される。

解析の手順は以下のようになる。

1. Webページの表示されている部分のスクリーンショットを取得する。
2. 各ピクセルの色をL*a*b*色空間上でもっとも近いWebセーフカラーにマッピングし、RGB各成分6階調、216色に減色する。
3. 減色されたスクリーンショットからページ全体のカラーシグネチャと32pixel×32pixelのブロックごとの部分カラーシグネチャを作成、特徴量として蓄積する。なお、計算量の削減のため、ブロックに占める画素の割合が1%に満たない色は除く。
4. 3.と同時に、ページ内で出現頻度の高かった色をキーとする逆引きインデクスにページを登録する。

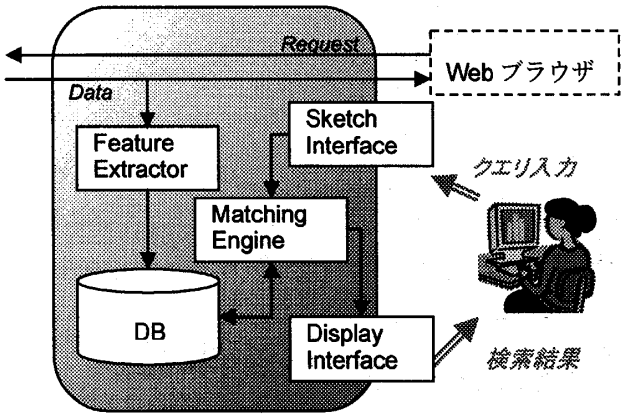


図1. 提案システムの構成

このインデクスには、色のpixel数も合わせて登録されており、pixel数でソート済みである。

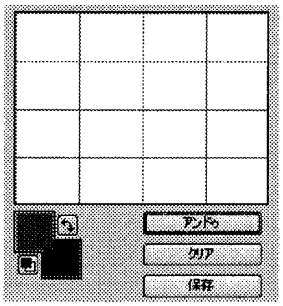
3. 類似度判定アルゴリズム

3.1 ユーザの入力

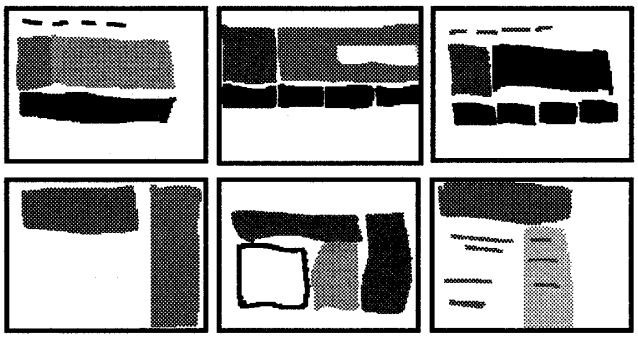
類似度判定アルゴリズムの設計に先立ち、図2(a)に示したスケッチ入力インターフェースを試作、これを用いて実際のWebページを見ながらスケッチしてもらう実験を行った。試作したインターフェースはマウスによるフリーハンド入力のほか、閉領域の塗りつぶし、アンドゥ機能を備える。同様の実験は先行研究[2]でも行われていたが、本実験とはモノクロ入力のインターフェースを用いている点、比較的単純なデザインのページを対象とし、レイアウトに特化した分析を行っている点が異なる。

実験には大学院生7名が参加した。いずれも日常的にWebを閲覧しており、Webページの作成経験を持っている。被験者の入力したスケッチの例を図2(b)に示す。実験では入力の順序も観察したが、すべての被験者に共通して、周囲と異なる背景色・背景画像を持つ要素を適当に入力し、その後テキスト、画像などのコンテンツの入力に移るという振る舞いが見られた。実験後のアンケートでも、「スケッチ入力式のインターフェースではコンテンツ画像は表現が困難なため、入力は避けたい」「代わりにアイコンのようなもので済ませられると良い」という意見が多く寄せられた。

そのほかに興味深い点として、多くの被験者で入力の際に要素同士の重なりを避け、画面を塗り分けるように入力する様子が観察された。このような入力スキームはDocument Object Model (DOM)の基本となるツリー状、入れ子上の構造と相性が悪く、DOM解析だけでは視覚的特徴の分析が困難であることを示している。



(a)



(b)

図 2. (a)スケッチ入力インターフェース; (b) ユーザの入力例

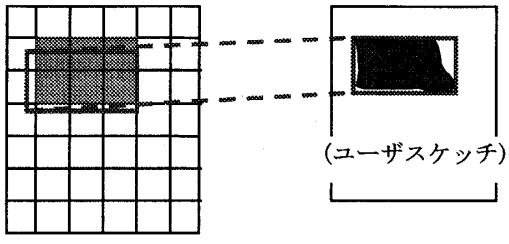
3.2 マッチング手法

マッチングはスケッチ入力インターフェースが一つの操作（色の変更、線の描画、または領域の塗りつぶし）を受け取るごとに行う。なお、試作インターフェースが備えていたアンドゥ機能は未実装である。

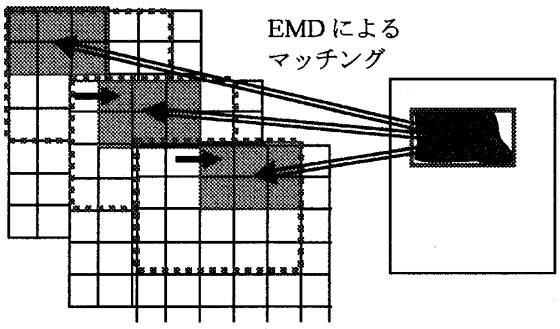
まず、ユーザが描画に使用する色を選択した段階で、選択されている色と類似する色を含むページを優先マッチング候補とする絞込みを行う。

一回のスケッチ入力に対するマッチングに際しては、ユーザの入力したスケッチから入力により変化の生じた部分を囲む矩形を抽出、この部分から生成したL*a*b*色空間でのカラーシグネチャをクエリQの特徴量S_Qとして用いる。マッチングの手順は以下のようになる。

1. クエリ領域をデータベース中の Web ページ P において相対位置、相対サイズが最も近いブロック群 Pi に対応付ける。このブロック群とのマッチングを考える。(図 3(a))
2. あらかじめ抽出しておいたブロック単位のカラーシグネチャを合成し、Piの特徴量S_{Pi}を算出する。
3. EMD を用いて Q と Pi との類似度を算出する。
4. Pi を上下左右に 1 ブロックずらしたもの、それぞれに対しても 1. – 3.の手順により類似度を算出する。(図 3 (b))
5. 1. – 4.で算出された類似度の最大値と最小値の差をページとクエリとの類似度とする。
   なお、クエリ領域に対応するブロックがページ全体に及び、4 の手順が省略された場合には、手順 3 で算出された値を類似度とする。



(a)



(b)

図 3. マッチング手法： (a) ブロックへの対応付け; (b) ページ内探索マッチング

最後に、それまでの入力に対して算出された類似度に減衰定数 c (< 1)を乗じたものに上記の値を加え、クエリに対する類似度とする。入力ごとに順位が大きく変動しないように、クエリ全体の類似度はこのような算出方法としているが、状況によっては、まったく異なる結果が望ましい場合もあり、ユーザの操作に応じて処理を変えていくなどの改良が考えられる。今後、本手法の有効性を示すための実験とあわせて検討を行う予定である。

4. おわりに

本稿では、利用者の入力したスケッチに基づき、色やレイアウトなどのデザインの近い Web ページを検索する手法を提案した。Web ページの見た目の特性に加え、ユーザの入力の特性を最大限に活用し、部分的な情報を元に段階的な絞込みを行うことで簡易なアルゴリズムで照合を行うことが可能である。

プロトタイプで採用した検索手法は照合時にクエリに合わせて特徴量を再構成するコストがかかるため、検索の高速化には限界がある。本システムにおいては Web 閲覧時のアイドル時間が利用できるため、バックグラウンドでの特徴量抽出を強化し、簡便な類似度判定アルゴリズムを採用することも考えられる。今後、本手法の改良と並行して取り組んでゆく。

参考文献

[1] Y. Rubner, et al., "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval", *Proc. of the ARPA Image Understanding Workshop*, pp. 661-668 (1997)

[2] 橋本泰成, 五十嵐健夫, 「レイアウトによる Web ページ検索」, インタラクション 2004 , pp.113-120(2004)