

登録番号 B4-8

WWW に対するマルチメディアデータ検索エンジンの HTML 構文を活かしたスコア付け手法の提案

杉尾 敏康[†] 竹野 浩^{††} 藤本 典幸[†] 萩原 兼一[†]

A Scoring Method for a Multimedia Data Search Engine on WWW

Toshiyasu SUGIO[†], Hiroshi TAKENO^{††}, Noriyuki FUJIMOTO[†], and Kenichi HAGIHARA[†]

あらまし WWW には画像や音楽などのマルチメディアデータが膨大に存在する。そのため、ユーザは自分の希望するマルチメディアデータを効率良く検索することが難しい。そこで、WWW に存在するマルチメディアデータを効率良く検索し、ユーザに提供するマルチメディアデータ検索エンジンが研究されている。マルチメディアデータ検索エンジンは、スコアによってマルチメディアデータと単語とを関連付けし、スコアの高い順に検索結果を返す。本研究では、HTML の構文構造を活かしたスコア付け手法（構文スコア）を提案し、HTML のタグを利用した既存のスコア付け手法（周辺タグスコア）と組み合わせることによって、検索結果における適合率の向上を目指した。そして、27 種類の質問を用いて提案手法を評価した結果、周辺タグスコアよりも適合率が向上することがわかった。また、マルチメディアデータ検索エンジンの 1 つである Google の画像検索エンジンと本システムの比較実験の結果、本システムにより Google 画像検索の適合率および再現率を向上できることがわかった。提案手法を用いれば、マルチメディアデータ検索エンジンにおける適合率および再現率を向上でき、ユーザの希望したマルチメディアデータを迅速に提供できると考える。

キーワード WWW, 検索エンジン, マルチメディアデータ, 文書解析, HTML

1. はじめに

現在インターネットの急速な発展により、膨大な数の画像や音楽などの **マルチメディアデータ (Multimedia Data, 以下, MD)** が World Wide Web (WWW) に存在する。この MD を WWW から検索するシステムとして、画像検索エンジンや音楽検索エンジンが存在する [1][2][3]。本研究では、これらのシステムを総称して **マルチメディアデータ検索エンジン (Multimedia Data Search Engine, 以下, MDSE)** と呼ぶ。ユーザは MDSE に質問を入力することによって、WWW に数多く存在する MD からユーザの希望する MD を検索することができる。一般に、MDSE は検索手法の違いによって、以下の 2

種類に分類できる。

(1) 文書解析に基づく検索

文書検索と同様に、ユーザは検索したい MD に関連がある単語を質問とする。MDSE は MD を含む HTML 文書内の出現単語によって MD を検索する。文書内に MD が存在すれば、MD の注釈やファイル名、および MD の周辺に出現する単語と MD の関連の強さをスコア付けする。

(2) 内容解析に基づく検索

ユーザはシステムがあらかじめ提供するサンプル画像や、検索したい MD の色や形などを質問とする。MDSE は MD のデータ内容を解析し、MD の特徴量によって検索を行う。一般に内容解析では、高次元の特徴量を扱う必要があるため、メモリ資源や前処理の負担が大きく、検索処理が複雑になる (e.g., [4][5])。

本研究は、WWW に対する MDSE を対象としているため、検索エンジンが扱う文書は HTML で記述されていると想定する。HTML で記述された文書は、

[†] 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University
^{††} NTT サイバーソリューション研究所
NTT CyberSolutions Laboratories

MD に対する注釈・説明と見なせるため、この文書を解析すれば MD の内容を解析しなくとも適合率の高い検索が行えると考えられる。また、MD の内容を解析するためには、MD の種類に応じて解析法や質問を用意する必要がある。しかし、文書解析では HTML の内容を解析するため、MD の種類によらず同一の手法で解析することができる。以上より本研究では、文書解析に基づく検索手法に着目する。

文書解析に基づくスコア付け手法の 1 つに、HTML のタグに基づいたスコア付け（以下、**周辺タグスコア**）がある [6]。周辺タグスコアは MD と関連が強いタグにあらかじめ大きい重みを割り振っておき、タグ内に出現する単語と MD の関連の強さを、タグに割り振られた重みによってスコア付けする。また、MD の周辺に出現する単語にも高い重みを割り振り、MD との関連を強調する。本研究では、周辺タグスコアに HTML 文書の構文構造を考慮したスコア付け（以下、**構文スコア**）を組み合わせてることによって、検索結果における適合率および再現率の向上を目指した。そして、27 種類の質問を用いて提案手法を評価した結果、周辺タグスコアよりも適合率が向上することがわかった。さらに、MDSE の 1 つである Google の画像検索エンジン [1] との比較実験の結果、本システムにより Google 画像検索の適合率および再現率を向上させることがわかった。提案手法を用いてスコア付けを行えば、MDSE の適合率および再現率が向上し、ユーザは希望した MD を迅速に取得できると考える。

2. マルチメディアデータ検索エンジン (MDSE)

MDSE は、WWW から文書を収集し、文書内に存在する MD m と単語 w のすべてのペアに対するスコア $S(m, w)$ の集合とともに、収集した文書をデータベースに保持している。ユーザは通常の全文検索エンジンの場合と同様に、キーワードと and, or, not などの演算子からなる質問を用いて MD を検索できる。MDSE は $S(m, w)$ に基づいて質問に適合する MD を検索し、スコアの降順に並べた MD を検索結果としてユーザに返す。

MDSE の検索精度を評価する値として、以下の式で表される**適合率**と**再現率**がある。適合率は検索結果のうち質問に適合している MD の数（以下、True_MD）の割合を表し、適合率が高ければ、質問に適合する MD が検索結果に多いことを表す。再現率は収集した

文書集合に存在するすべての MD（以下、all_MD）のうち、質問に適合する MD を洩れなく検索できている割合を表し、その値が高ければ all_MD から質問に適合する MD を洩れなく検索できていることを表す。

$$\text{適合率} = \frac{\text{True_MD}}{\text{検索結果に現れた MD 数}}$$

$$\text{再現率} = \frac{\text{True_MD}}{\text{質問に適合する MD 数} \in \text{all_MD}}$$

一般に適合率と再現率にはトレードオフが存在する [7]。MDSE は適合率と再現率がともに高いほど検索精度が高いと見なされる。適合率と再現率は MDSE が使用するスコア付け手法に依存するため、スコア付け手法が適切でなければ MDSE の検索精度は低下する。

本研究で開発している MDSE は、検索ゲートウェイと検索サーバから構成される（図 1）。ユーザは検索ゲートウェイ上に実装された cgi 画面から、検索したい MD と関連がある質問を入力することによって、検索結果を取得できる。本システムは以下の MD に対する検索を提供している。ユーザは検索の際、検索する MD の種類を 1 つだけ指定する。

- 静止画像 (JPEG, GIF など)
- 動画像 (MPEG, AVI など)
- 音声 (MP3, WAV など)
- 文書データ (PDF, PS など)

本システムの検索結果画面のサンプルを図 2 に示す。図 2 (上) は検索単語「桜」で JPEG 画像を検索した結果である。本システムは検索結果が画像データの場合はサムネイル表示を行う。図 2 (下) は検索単語「MPEG」で PDF 文書を検索した結果であり、表示されたアイコンをクリックすると文書の内容が表示される。また、検索結果の MD が存在する文書のタイトルや URL を表示し、文書へのアクセス手段を提供する。他にも音声や動画データの検索も提供する。

3. 周辺タグスコア

MD と単語のペアに対するスコア付け手法の 1 つに HTML のタグを利用した周辺タグスコア [6] がある。本節では周辺タグスコアおよびその問題点について説明する。

3.1 周辺タグスコアによるスコア付け手法

HTML 文書内に出現する単語は、その文書内に存在する MD と関連がある可能性が高い。しかし、存

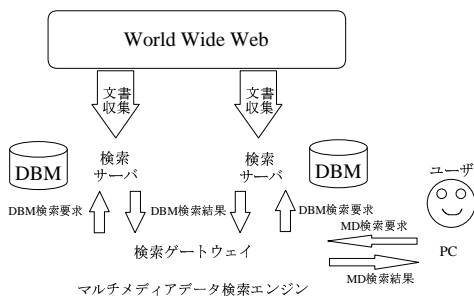


図 1 マルチメディアデータ検索エンジン (MDSE)
Fig.1 A Multimedia Data Search Engine(MDSE)

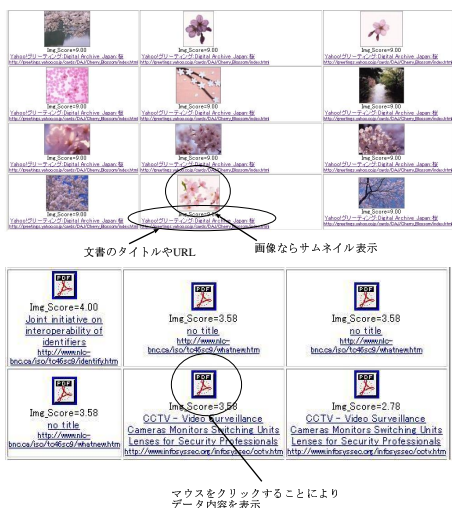


図 2 検索結果画面のサンプル (上図:「桜」で JPEG 画像を検索した結果。下図:「MPEG」で PDF 文書を検索した結果)

Fig. 2 Examples of display of retrieved results

在する MD はすべての単語に等しく関連があるとは限らない。そこで、周辺タグスコアは HTML 文書がタグで属性化されていることに着目し、単語を含むタグの種類によりスコアを変える。すなわち周辺タグスコアでは、MD と関連が強いと考えられるタグおよび属性に高いタグの重みを割り振り、タグ内に出現する単語と MD のペアにタグの重みに従ってスコアを与える。これにより、関連の強い MD と単語のペアに対して高いスコアを与えることができる。表 1 にタグの重みを示す。MD のファイル名を表す IMG タグ内の SRC 属性, MD の注釈を表す ALT 属性, 文の主題を表す TITLE タグという順に高いスコアが割り当てられている。表 1 を用いてタグの重みを割り振った例を図 3 に示す。単語と MD のペアには、

表 1 タグスコアによるタグの重み
Table 1 Weights of Tag

HTML タグまたは属性	重み
SRC	8.00
ALT field of IMG	6.00
TITLE	5.00
H1	4.00
H2	3.60
H3	3.35
H4	2.40
H5	2.30
H6	2.20
B	3.00
EM	2.70
I	2.70
STRONG	2.50
<No Tag>	1.00

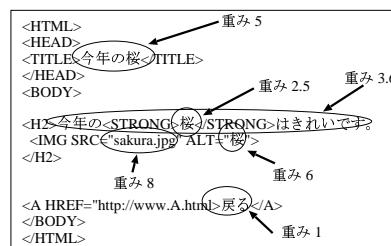


図 3 タグスコアによるスコア付けの例
Fig. 3 An example of scoring by Tag_score

単語が所属するタグの重みを用いてスコアを付ける (単語「桜」と MD (sakura.jpg) のペアにはスコア $S(\text{sakura.jpg}, \text{「桜」}) = 17.1 (=5+3.6+2.5+6)$ が与えられる)。

また、周辺タグスコアはタグの重みに加えて MD の周辺に出現する単語に高い重みを割り振る。周辺タグスコアによって、HTML 文書内の単語と MD のペアに対してスコアを与える方法を以下に示す。

- 入力: HTML 文書 H
出力: H 内に存在する MD m と単語 w の全てのペアに対する $S(m, w)$ の集合
- (1) foreach 単語 $w \in H$ do
 - (1.1) $S'(w) = 0$
 - (1.2) foreach 表 1 のタグ $t \in H$ do
 - (1.2.1) if t が w を含む then $S'(w) +=$ 表 1 による t の重み
 - (2) foreach MD $m \in H$ do
 - (2.1) $S(m, w) = S'(w)$
 - (2.1.1) foreach $w \in H$ do
 - (2.1.1.1) if $0 < \text{pos}(m, w) < \text{b_dist}$ then $S(m, w) += \rho \times e^{-2.0 \times \text{pos}(m, w) / \text{b_dist}}$
 - (2.1.1.2) else if $\text{a_dist} < \text{pos}(m, w) < 0$ then $S(m, w) += \rho \times e^{-2.0 \times \text{pos}(m, w) / \text{a_dist}}$
- (1) から (2.1) まではタグによる重み付けを表す。

$pos(m, w)$ は MD m と単語 w の距離を表しており、 m より前に出現する単語との距離を正の数、後に出現する単語との距離を負の数で表す (m より 1 つ前に出現する単語 w は $pos(m, w)=1$, 1 つ後に出現する単語 w' は $pos(m, w')=-1$ となる). b_dist , a_dist の値はそれぞれ 10 と -20 であり、MD より前に出現する 10 単語、後に出現する 20 単語に周辺タグスコアによる重みを加算する ((2.1.1.1), (2.1.1.2)). また、 ρ の値は 5.0 であり、MD に最も近い単語は TITLE タグ内の単語に割り振られるタグの重みに近い値となる [6]. 周辺タグスコアにおいて重みを与える式 $\rho \times e^{-2.0 \times pos(m, w)/dist}$ は、以下の仮定の元に得られる.

- HTML 文書内において MD m の周辺に出現する単語 w は m と関連がある.
- w と m の関連の強さは m からの距離によって指数的に低下する.

上記の仮定は常に成り立つとは限らないが、多数の HTML 文書の解析の結果、適切であることがわかっている [6].

周辺タグスコアでは文書内に出現回数が多い単語ほど重みが増加する. そこで、出現回数が多い単語の重みと少ない単語の重みの差を抑えるために、文書集合における単語 w の貴重度 $G(w)$ [6] をスコア付けに導入し、 $G(w)$ を用いてスコアを重み付けする. $G(w)$ を求める手法は様々提案されているが、文献 [8] に基づく性能評価より、log-entropy 法を用いる. $G(w)$ を導入してスコア $S(m, w)$ を求める式を以下に示す.

$$S'(m, w) = L(m, w) \times G(w)$$

$$L(m, w) = \log(S(m, w) + 1)$$

$$G(w) = 1 - \sum_k \frac{p_{wk} \log(p_{wk})}{\log(ndocs)}$$

$$p_{wk} = tf_{wk} / \sum_k tf_{wk}$$

tf_{wk} は文書 k における単語 w の出現回数を表し、 $ndocs$ は文書集合の全体の文書数を表す. 周辺タグスコアでは以上の式で求まる $S'(m, w)$ を m と w のペアに対するスコアとし、検索に使用する.

3.2 周辺タグスコアの問題点

周辺タグスコアでは以下の問題点 P1, P2 が発生する.

P1 MD m と距離が遠い単語 w でも $S(m, w)$ は 1 以上となる. そのため、図 4 のように MD (sakura.jpg) と関連の弱い単語 (「犬」) も、その出現回数が多くなると $S(sakura.jpg, 「犬」)$ が上昇

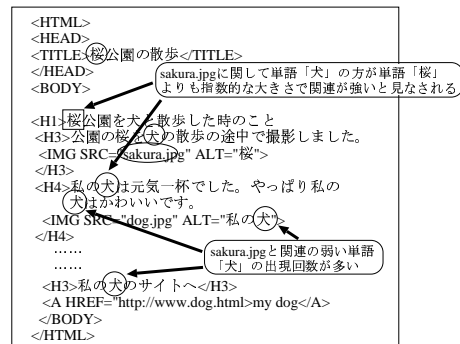


図 4 周辺タグスコアの問題点
Fig. 4 Problem of Around-Tag-score

し、関連が強いと見なされる. その結果、検索結果の適合率が低下する (「犬」で検索した際に sakura.jpg が検索結果の上位にくる可能性が生じる).

P2 図 4 の場合、H4 タグ内の単語「犬」は H1 タグ内の単語「桜」よりも sakura.jpg とソースファイル内の距離が近い. そのため、 $S(sakura.jpg, 「犬」)$ は $S(sakura.jpg, 「桜」)$ よりも指数的に大きい値となり、単語「犬」の方が sakura.jpg と関連が強いと見なされる. その結果、「犬」で検索を行った際に sakura.jpg が検索結果の上位に出現する可能性があり、検索結果の適合率が低下する.

4. 提案するスコア付け手法

問題点 P1, P2 より、周辺タグスコアは扱う文書によっては検索結果の適合率が低下する. そこで本研究では、HTML 文書を木構造と解釈し、木構造における距離を考慮したスコア付け手法 (以下、構文スコア) を提案する. 構文スコアを用いれば、文書内の MD と木構造において距離に近い単語のペアに対するスコア付けを強調でき、検索結果の適合率を向上できる. また本研究では、周辺タグスコアと構文スコアを組み合わせたスコア付け手法 (以下、周辺構文タグスコア) を提案する. 周辺構文タグスコアは、MD m と文書構文において距離に近い単語や、 m と関連が強いタグに含まれる単語および m の周辺に出現する単語を強調する. 周辺構文タグスコアは、周辺タグスコアによって生じる問題点 P1, P2 を軽減でき、検索結果の適合率をさらに向上できる. 以下まず、構文スコアおよび周辺構文タグスコアで使用する HTML 木について述べ、構文スコアによるスコアの付け方および、周辺構文タグスコアについて述べる.

4.1 HTML 木

HTML 文書はタグの階層構造で記述するため、木構造と見なせる。本研究は、この木構造を HTML 木と呼ぶ。HTML 木の作成法を以下に示す。

- (1) HTML のタグを木のノードとする。
- (2) タグの階層構造に従ってノード間をアークで結ぶ。
- (3) タグ t 内に文が存在すれば t の子ノードとして新たにノード (以下、text ノード) を作成する。

図 3 の HTML 文書から作成した HTML 木を図 5 に示す。HTML 木において、MD を持つノードをその MD の保持者と呼ぶ。図 5 では MD (sakura.jpg) を持つ IMG タグが保持者である。保持者には兄弟のノードや親のノードが存在する場合がある。本論文では兄弟を以下のように区別する。文書内の出現順序が早いノードを兄、遅いノードを弟とし、兄弟の中で最初に文書内に出現するノードを 1 番目の兄、その次に出現するノードを 2 番目の兄、などとする。これらのノードと保持者との関係を HTML 木の近隣関係と呼ぶ (表 2)。表 2 で下に書いてあるノードほど、保持者との近隣関係が遠いと定義する。文書内の上位に位置する文は下位に位置する文よりも、MD の内容を反映する傾向がある [9] ため、弟よりも兄の方が保持者に近いと考える。図 5 では、IMG タグが保持者であるのに対して、text ノード (1) が 1 つ上の兄、H2 タグが親、そして A タグが親の弟となる。HTML 木では、次の性質 N1 が生じると想定する。

N1 保持者からの近隣関係が近いノードほど、保持者との関連が強い。

また、図 6 のように保持者からの近隣関係が同じノードでも、そのノードから保持者までの間に存在するノード数 (以下、通過ノード数) が異なる場合がある。図 6 の (a)(b) では、文書の先頭の H1 タグは両方ともに保持者 (IMG タグ) の 2 つ上の兄であるが、(b) の H1 タグは通過ノード数が多く、タグ内の内容は (a) の H1 タグに比べて保持者との関連が弱くなっている。よって HTML 木には、次の性質 N2 が生じると想定する。

N2 保持者からの近隣関係が同じであるノードは、通過ノード数が多くなるほど保持者との関連が弱くなる傾向がある。

本研究では、HTML 木の性質 N1, N2 を利用してスコア付けを行うことを考えた。

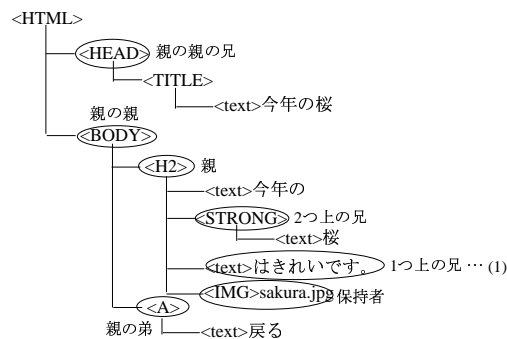


図 5 図 3 の HTML 文書の HTML 木
Fig. 5 The HTML tree of the HTML document in figure 3

表 2 構文スコアによるノードの重み
Table 2 Weight of node

近隣関係	ノードの重み
保持者	5.00
1 つ上の兄	4.00
1 つ下の弟	3.60
2 つ上の兄	3.35
2 つ下の弟	2.40
親	3.00
親の兄	2.30
親の弟	2.20
親の親	2.00
親の親の兄	1.80
親の親の弟	1.70
上記以外のノード	0.00

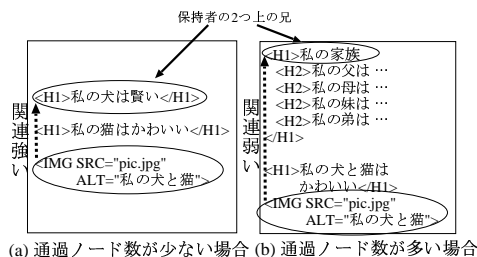


図 6 通過ノード数による保持者との関連性の変化
Fig. 6 Strength of relation between a node and a holder due to the number of passing nodes

4.2 構文スコア

構文スコアは HTML 文書を HTML 木として解析し、木の近隣関係に基づいてスコア付けを行う。構文スコアによるスコア付け手法の手順を以下に示す。

- 入力: HTML 文書 H
 出力: H 内に存在する MD m と単語 w の全ての組み合わせに対する $S(m, w)$ の集合
- (1) foreach MD $m \in H$ do
 - (1.1) foreach ノード $n \in H$ (根から深さ優先順に) do
 - (1.1.1) $N = |n \text{ のノード番号} - m \text{ の保持者のノード番号}|$

- 1 |

- (1.1.2) n の重み = 表 2 に基づく n の重み $- N \times \alpha$
- (2) foreach MD $m \in H$ do
- (2.1) foreach ノード $n \in H$ (根から深さ優先順に) do
- (2.1.1) if n の重み ≤ 0 then n の重み = n の親の重み $- N \times \alpha$
- (3) foreach text ノード $t \in H$ do
- (3.1) foreach 単語 $w \in t$ do
- (3.1.1) $S(m, w) = t$ の重み

表 2 のノードの重みは、現在のシステムでの実験的な値である。性質 N1 を考慮し、保持者からの近隣関係に近いほどノードの重みを高く設定している。また構文スコアでは、性質 N2 を考慮するために、重みを付けるノードから保持者までの通過ノード数を求め、スコア付けに反映する。通過ノード数は、根から深さ優先順につけた各ノードのノード番号と、MD を保持するノードのノード番号との差をとった値の絶対値から求める (上記 (1.1.1))。また現在、実験的に α の値を 0.02 に固定している。

図 5 の HTML 木に対して構文スコアを適用した例を図 7 に示す。図 7 から、構文スコアは保持者からの近隣関係に近いほど大きいノードの重みを与えることがわかる。

図 4 の文書に対して構文スコアを適用した場合を考える。図 4 の文書では、MD (sakura.jpg) と関連の弱い単語「犬」の出現回数が多いが、文書構文において MD との距離が遠いため、単語「犬」と MD のペアに対するスコアは出現回数に依存して上昇しない。その結果、「犬」で検索した際に、sakura.jpg が検索結果の上位にくる可能性が低下し、周辺タグスコアの問題点 P1 が軽減されると考えられる。また、構文スコアでは近隣関係に近い単語と MD に対する関連を強調するため、図 4 の H4 タグ内の単語「犬」より、H1 タグ内の単語「桜」の方が sakura.jpg と関連が強いと見なされる。これにより、周辺タグスコアの問題点 P2 が軽減される。

以上より、構文スコアを用いれば、検索結果の適合率が向上する。しかし、構文スコアのみでは、MD と近隣関係に近い単語と MD のペアに対するスコアのみが上昇し、MD と関連が強いタグ内の単語と MD のペアに対するスコアが低下する可能性がある (図 7 では、text ノード (1) の $S(\text{sakura.jpg}, \text{「桜」})$ よりも、text ノード (2) の $S(\text{sakura.jpg}, \text{「戻る」})$ の方が高くなるため、「戻る」で検索を行った場合に、sakura.jpg が検索結果の上位に現れる可能性が生じる)。また、構

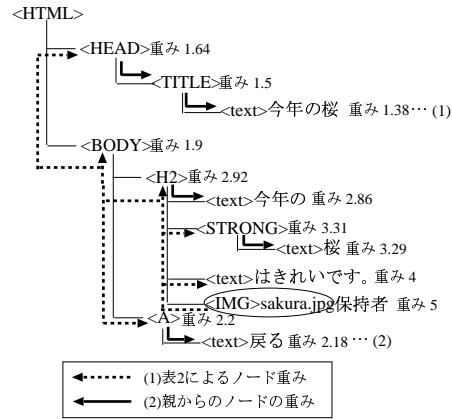


図 7 HTML 木におけるスコア付け
Fig. 7 Scoring the HTML tree in figure 3

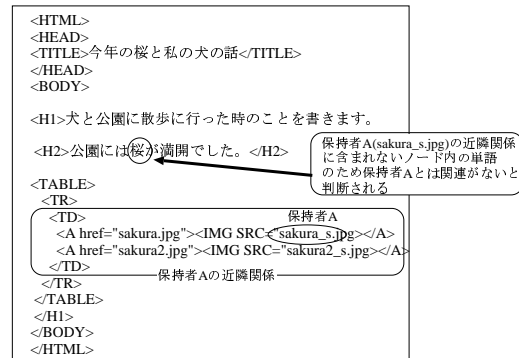


図 8 構文スコアの問題点
Fig. 8 Problem of Construct-score

文スコアは MD m (保持者) の近隣関係に text ノードが出現しない場合、 m と関連がある単語 w と m のペアにスコアを割り振れない。図 8 では、保持者 A (sakura_s.jpg) と関連がある H2 タグ内の単語「桜」は、保持者 A の近隣関係に含まれないタグ内に存在するため、保持者 A とは関連がないと見なされる。その結果、保持者 A は単語「桜」で検索を行った場合に検索結果に現れず、MDSE の適合率および再現率が低下すると考えられる。以上の場合を考慮するために、構文スコアに周辺タグスコアを組み合わせる必要があると考えられる。

4.3 周辺構文タグスコア

周辺構文タグスコアでは、文書内の MD に対して構文スコアと周辺タグスコアの両方を適用する。これにより、MD と近隣関係が近く、さらに MD と関連が強いタグ内の単語と MD のペアに対するスコア付

けを強調できる。周辺構文タグスコアは以下の手順に基づいてスコア付けを行う。

(1) タグスコアによって文書内の MD m と単語 w のペアに対するスコア $\text{Tag}(m, w)$ を求める。

(2) 構文スコアによって、MD m と近隣関係が近い単語 w のペアに対するスコア $\text{Con}(m, w)$ を求める。

(3) MD の周辺に出現する単語 w と MD m のペアに対するスコア $\text{Ard}(m, w)$ を求める。

b_dist , a_dist , ρ の値は周辺タグスコアで用いる値 ($b_dist=10$, $a_dist=-20$, $\rho=5.0$) を使用し、MD より前に出現する 10 単語、後に出現する 20 単語に重みを加算する。

(4) 以下の式によって求められる値を、 $S(m, w)$ として DBM に登録する。

$$S(m, w) = (\text{Con}(m, w) + \text{Ard}(m, w)) \times \text{Tag}(m, w)$$

MD m と近隣関係が遠く、 m の周辺に出現しない単語 w を m と関連付けないために、 $\text{Con}(m, w)=0$ かつ $\text{Ard}(m, w)=0$ であれば $S(m, w)$ を 0 とする。そして $S(m, w)$ が 0 の単語と MD のペアは DBM に登録しない。

また、周辺構文タグスコアでは貴重度 $G(w)$ をスコア付けに導入する。周辺構文タグスコアを用いれば、問題点 P1, P2 を軽減でき、検索結果の適合率および再現率を向上できる。

5. 評価

5.1 周辺タグスコアと周辺構文タグスコアの比較

本研究で提案したスコア付け手法の精度を評価するために、周辺タグスコアおよび周辺構文タグスコアを用いた MDSE を実装し、比較実験を行った。比較には、以下の式で定義する検索結果の到達率を用いた。

$$\text{到達率} = \frac{X \text{ 件のうち質問に適合する MD 数}}{\text{検索上位 } X \text{ 件}}$$

到達率は検索結果上位 X 件において質問に適合する MD の数の割合を表し、到達率が高ければスコア上位に質問に適合する MD が多いことを表す。本実験では X の値を 10 から 50 まで 10 刻みで変化させ、27 種類の質問を用いて静止画像を検索した場合の検索結果上位 X 件における到達率を比較した。周辺タグスコアおよび周辺構文タグスコアの 27 種類の質問に対する到達率の平均を図 9 に示す。図 9 から、周辺構文タグスコアを用いれば周辺タグスコアよりも検索結

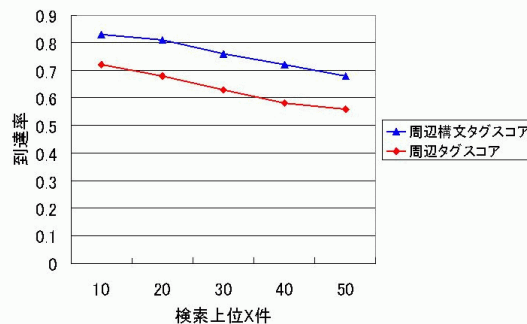


図 9 各スコア付け手法による到達率の平均値
Fig.9 Average of attainment for each methods

果の上位に質問に適合する MD を多く集められることがわかる。

上記の結果となる要因を以下に説明する。周辺構文タグスコアは、MD と近隣関係が近い単語と MD のペアに対するスコア付けおよび、MD と関連が強いタグ内の単語と MD のペアに対するスコア付けの両方を強調する。そして、近隣関係が遠い単語と MD のペアに対するスコア付けを削除する。これにより、周辺タグスコアの際に生じる問題点 P1, P2 を軽減でき、MD と関連の強い単語と MD のペアに対するスコアを向上できる。その結果、周辺タグスコアよりも検索結果の到達率が向上したと考えられる。

5.2 本システムを Google 画像検索エンジンの後処理に用いた場合の評価

MDSE の一種である Google の画像検索エンジン [1] と、周辺構文タグスコアを実装した本システムを Google の後処理に用いた場合の適合率および再現率の比較を行った (表 3)。本システムと Google の保持するデータベースは異なるため、比較実験では以下の手順に従って、本システムと Google が収集した文書集合を一致させた。

(1) 比較に使用する質問を用いて Google で画像検索し、検索結果の上位 20 件の画像が含まれる文書集合 D を取得する。

(2) D に含まれる MD をすべて人間の目で確認し、質問に適合する MD の数を数える (再現率の分母の値を求める)。

(3) (1) と同じ質問を用いて本システムおよび Google による画像検索を行い、検索結果の適合率お

よび再現率を計算する。なおこの際、 D に含まれない画像が検索結果に出現する場合があるが、それらの画像は適合率、再現率の計算に含めない。

表 3 から以下のことがわかる。

- 適合率は 27 種類の質問のうち、12 種類の質問に対して 0.01~0.17 向上しているが、15 種類に対しては 0.02~0.16 低下している。

- 再現率はすべての質問に対して 0.03~0.61 上昇している。

以上より、本システムは Google の画像検索エンジンで得られる検索結果に対して、適合率を最大で 0.17、再現率を最大で 0.61 向上できている。しかし、適合率は最大で 0.16 低下する場合がある。この適合率の低下を軽減するために、次のことを行う。

周辺構文タグスコアでは、単語 w と MD m のスコア $S(m, w)$ が 0 より大きければ、 m を質問に適合する MD として検索結果に出力する。そのため、検索結果の下位に質問に適合しない MD が集中して出現している可能性が高い。そこで、スコアの閾値 β を設定し、 β より $S(m, w)$ が高ければ、検索結果として m を出力する。

β の値を 1 から 4 まで 1 刻みに変化させた場合の適合率および再現率の比較結果を表 4 に示す。また、Google と比較して適合率および再現率が向上、低下した質問数を表 5 に示す。表 4、表 5 から以下のことがわかる。

- β の値が高くなると、Google よりも適合率が向上する質問が増加するが、再現率が低下する質問も出現する。

- $\beta = 3$ の場合に適合率が向上する質問数と再現率が向上する質問数のバランスがとれている。

以上より本システムは、 $\beta = 3$ の場合に Google の適合率および再現率を向上できる可能性が高い。また、 β の値を変化させることによって適合率および再現率の大きさは調整できる。そのため、検索時にユーザが β の値を指定すれば、適合率と再現率の優先度を調整でき、ユーザの意図に合った検索を提供できると考える。

6. まとめと今後の課題

本研究では、MDSE におけるスコア付け手法として、HTML 文書の構文構造を利用した構文スコアを提案した。そして、既存のスコア付け手法である周辺タグスコアと構文タグスコアを組み合わせた周辺構文

表 3 Google+本システムと Google 画像検索との適合率/再現率の比較

Table 3 Comparison of precision / recall between Google + our system and Google image search

質問	Google+本システム	Google
パソコン	0.81/0.53	0.64/0.25
桜	0.76/0.92	0.61/0.58
サッカー	0.81/0.65	0.68/0.58
イチロー	0.83/0.77	0.71/0.33
正月	0.74/0.74	0.65/0.28
結婚	0.78/0.78	0.71/0.55
犬	0.77/0.63	0.70/0.25
猫	0.76/0.59	0.71/0.41
平和	0.79/0.87	0.75/0.46
卒業式	0.89/0.60	0.86/0.25
熱帯魚	0.83/0.81	0.85/0.50
ボルシェ	0.87/0.89	0.91/0.57
USA	0.88/0.77	0.92/0.73
時計	0.78/0.91	0.84/0.48
干支	0.71/0.89	0.78/0.58
携帯電話	0.72/0.82	0.79/0.43
クリスマス	0.68/0.98	0.77/0.32
落葉	0.68/0.95	0.77/0.45
スニーカー	0.70/0.66	0.80/0.21
誕生日	0.51/0.96	0.63/0.54
日の出	0.65/0.94	0.78/0.45
富士山	0.64/0.88	0.80/0.62
秋 & 魚	0.71/0.73	0.64/0.32
黄色い & 花	0.69/0.72	0.68/0.27
田舎 & 景色	0.72/0.74	0.75/0.33
冬 & 日本海	0.75/0.92	0.82/0.38
美しい & 花火	0.58/0.81	0.73/0.50

タグスコアを提案し、周辺タグスコアとの比較実験の結果、検索結果の到達率が向上することがわかった。また、Google の画像検索との比較実験を行い、本システムにより Google 画像検索の適合率および再現率を向上できることがわかった。以上より、本システムは WWW から取得した MD のうち質問に適合する MD を迅速にユーザに提供できると考える。今後の課題として、ユーザにより良い MD を提供するために、MD の質を考慮したスコア付けを導入する必要がある。周辺構文タグスコアは MD の質をスコア付けに反映していないため、検索結果の上位に画質や音質の悪い MD が出現する可能性がある。

謝辞

本研究の一部は、日本学術振興会未来開拓学術研究推進事業 (JSPS-RFTF99I00903) および栢森情報科学振興財団の補助による。丁寧な査読をして下さった査読者に深く感謝する。

文 献

- [1] Google Inc. Google image search engine.
<http://images.google.com/>.

表 4 閾値 β を変化させた場合の適合率/再現率の比較
 Table 4 Comparison of precision / recall for each β between Google + our system and Google image search

質問	Google+本システム				Google
	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$	
パソコン	0.81/0.53	0.81/0.53	0.83/0.53	0.86/0.44	0.64/0.25
桜	0.76/0.92	0.88/0.88	0.90/0.69	0.94/0.63	0.61/0.58
サッカー	0.81/0.65	0.80/0.62	0.79/0.58	0.79/0.58	0.68/0.58
イチロー	0.85/0.77	0.88/0.77	0.89/0.77	0.90/0.68	0.71/0.33
正月	0.75/0.74	0.78/0.72	0.87/0.62	0.93/0.49	0.65/0.28
結婚	0.82/0.78	0.81/0.75	0.83/0.75	0.82/0.70	0.71/0.55
犬	0.80/0.63	0.82/0.63	0.83/0.53	0.85/0.45	0.70/0.25
猫	0.76/0.51	0.78/0.49	0.77/0.46	0.76/0.35	0.71/0.41
平和	0.79/0.87	0.75/0.85	0.78/0.79	0.77/0.69	0.75/0.46
卒業式	0.91/0.60	0.96/0.60	0.99/0.55	0.98/0.34	0.86/0.25
熱帯魚	0.82/0.78	0.88/0.78	0.88/0.64	0.94/0.42	0.85/0.50
ボルシェ	0.87/0.89	0.88/0.78	0.93/0.68	0.92/0.59	0.91/0.57
USA	0.88/0.77	0.88/0.77	0.88/0.77	0.87/0.67	0.92/0.73
時計	0.78/0.91	0.78/0.91	0.80/0.89	0.81/0.66	0.84/0.48
干支	0.71/0.89	0.78/0.86	0.78/0.81	0.80/0.56	0.78/0.58
携帯電話	0.72/0.82	0.75/0.82	0.80/0.82	0.62/0.52	0.79/0.43
クリスマス	0.69/0.96	0.72/0.89	0.77/0.64	0.93/0.47	0.77/0.32
落葉	0.69/0.95	0.76/0.89	0.81/0.89	0.79/0.82	0.77/0.45
スニーカー	0.72/0.66	0.76/0.64	0.75/0.22	0.84/0.17	0.80/0.21
誕生日	0.52/0.96	0.55/0.96	0.59/0.93	0.62/0.57	0.63/0.54
日の出	0.69/0.94	0.77/0.94	0.81/0.81	0.82/0.68	0.78/0.45
富士山	0.64/0.88	0.67/0.85	0.73/0.85	0.86/0.73	0.80/0.62
秋 & 魚	0.71/0.68	0.71/0.57	0.75/0.48	0.68/0.34	0.64/0.32
黄色い & 花	0.67/0.50	0.73/0.34	0.75/0.19	0.80/0.06	0.68/0.27
田舎 & 景色	0.75/0.72	0.77/0.43	0.88/0.15	1.00/0.07	0.75/0.33
冬 & 日本海	0.80/0.85	0.84/0.85	0.87/0.85	0.88/0.73	0.82/0.38
美しい & 花火	0.61/0.78	0.68/0.72	0.70/0.59	0.79/0.49	0.73/0.50

表 5 閾値 β を変化した場合に、Google と比べて適合率および再現率が向上、低下した質問数

Table 5 Relation between threshold β and precision/recall

閾値 β	適合率			再現率		
	向上	変化なし	低下	向上	変化なし	低下
$\beta = 1$	11	1	15	27	0	0
$\beta = 2$	14	2	11	27	0	0
$\beta = 3$	19	2	6	24	1	2
$\beta = 4$	23	0	4	18	1	8

- [2] Lycos Inc. Lycos internet guide.
http://multimedia.lycos.com/.
- [3] AltaVista Company. http://www.altavista.com/.
- [4] C. Frankel, M. Swain, and V. Athitsos: "Web-seer: An Image Search Engine for the World Wide Web", University of Chicago Technical Report TR-9614 (1996).
- [5] T. Gevers and A. Smeulders: "The PicToSeek WWW Image Search System", In Proc. IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, pages 264-269, June (1997).
- [6] M. La Cascia, S. Sethi, and S. Sclaroff: Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web, Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, pages 24-28 (1998).
- [7] S. Lawrence and C. L. Giles: "Searching the Web: General and Scientific Information Access", IEEE Communications, Vol.37, No.1, pages 116-122 (1999).
- [8] S. T. Dumais: "Improving the Retrieval of Information from External Sources", Behav. Res. Meth. Inst. Comput., Vol.23, No.2, pages 229-236 (1991).
- [9] R. Lempel and A. Soffer: "PicASHOW: Pictorial Authority Search by Hyperlinks on the Web", In Proc. of the WWW10 Conference, Hong-Kong, May (2001).