Webページ構造を考慮したキーワードによる画像の内容特定

大鹿　広憲　　山名　早人

早稲田大学理工学部情報学科

1　はじめに

近年のインターネットやデジタルカメラなどの普及により、WWW 上で扱われる画像の数は飛躍的に増えてきている。それに伴い、画像データベースが必要とされ、画像による内容の特定する必要性が出てきている。

HTML ファイル内でテキストによる画像の内容特定には<alt>タグを参照するのが最も有効的だと考えられるが、検索エンジンからランダムに抽出したページ 160 ページに対して<alt>タグが付加されているかどうかを調べたところ (表 1)、個人作成のページでは実際に画像に<alt>タグが付加されているものは半数以下である。(表 1)

表 1: alt タグ付加の割合 (数字は画像数)

	ある	ない	割合 (%)
個人用のページ	66	160	41.25
ニュース記事	11	4	73.33

2　研究の背景

<alt>タグがない場合、画像内容を特定する方法として、周辺のテキストを参照する方法が考えられる。しかし、「周辺のテキスト」の定義はあいまいで、関係のないテキストまでが抽出されてしまうことがある。

そこで本稿では、Web ページの構造に着目し、「WWW 上の画像を説明するキーワードはその画像ファイルの周辺に含まれていることが多い」という前提に基づいて、タグの構造を解析することによって近接テキストがどの程度がどこまでの範囲になるのかを検証する手法を提案する。

An Image Identification Scheme Using The Keyword Based on Web Page Structure
Hironori Oshika　Hayato Yamana
School of Department of Information and Computer Science, Science and Engineering, Waseda University

3　HTML 文書の解析

HTML 文書の解析に基づいた画像内容特定については既にいくつかの手法が提案されている。

- 画像リンクタグ、強調キーワードタグ、見出しタグ、画像表示タグの前後 3 行に着目し、画像ファイルとキーワードとの関連性を評価する方法 [1]
- 製品紹介のホームページを例に、ホームページ中に存在する画像内容を隣接するテキスト中に存在するキーワードと HTML タグの情報などのレイアウト情報からタグに優先順位をつけ、そのタグでレイアウトされたテキストを対象文章として抽出する手法 [2]
- 画像ファイルにキーワードを付加する方法として修飾されているタグによって重みつけを行う方法 [3]

しかし、[2] では製品紹介ページに限定しているので、適用範囲が狭く、[3] でも、ページの構造によっては、一定の重み付けだと十分な内容特定ができないと考えられる。

4　Web 構造の解析による近接範囲の判定

4.1　Web ページにおける画像の配置

実際、WWW 上の画像はどのように Web 上のページでどのように配置されているかを調べるために、ど

のようなタグで修飾されているかを wget でランダムに集めた 57 の Web ページを対象に調査を行った。その結果、約半数以上の画像が<table>タグで修飾されていることが分かった。(表 2)

表 2: Web ページにおける画像の位置

	画像数	割合 (%)
<table>タグ	42	77.8
<a>タグ	7	12.9
タグによる修飾なし	5	9.2

4.2 画像とテキストとの近接度の考察

画像の位置が<table>タグ内にある場合とない場合のとでは、画像と近接であると判断できるテキストの範囲が異なってくると考えられる。

- **画像が table タグで囲まれていない場合**
  タイトルや<font>や見出しタグなどで文字修飾されているテキストが画像内容を表している可能性が高い。また、テキストの情報が少ないときは、<meta>タグのキーワードも画像内容を表している可能性があると考えられる。

- **画像が table タグで囲まれている場合**
  表形式に画像が配置されていることが多いので、近接の範囲を狭める

4.3 タグの解析による近接度の判定

4.2 の考察を元にした画像ファイルにおける近接テキストを求めるためのアルゴリズムは以下のとおりである。

1. まず、対象の Web ページに対し、タグ解析を行う。
2. 画像ファイル、テキストそれぞれに対し、修飾されているタグを調べ、ハッシュに格納する。
3. 画像ファイルとテキストが隣どうしにあった場合の距離量を 1 とし、これを基準としたタグによる距離量を求める。
4. 画像ファイル数、画像ファイルの修飾タグにより、その画像ファイルに対する「近接範囲」を定める。(「距離量」はタグとテキストの頻度によって範囲を変える)
5. 「近接範囲」に含まれるテキストを抽出する

5 実験・考察

4.3 で提案したアルゴリズムをもとに、ランダムに抽出した Web ページを用いて、近接テキストの抽出の実験を行った。小さな画像を非対象とするために対象の画像の大きさを 65 × 65(pixels) 以上に指定した。抽出された近接テキストのうち、画像の内容と関係のないテキストが含まれていた場合を不正解とした。その結果、タグに基づいた近接範囲からのテキストを抽出した場合に、分類の精度が向上した。

表 3: 画像の近接テキストの抽出の実験結果

	正解数	不正解数	精度 (%)
提案手法	37	10	78.7 %
従来手法	23	24	48.9 %

6 今後の課題

本稿では、Web ページにおける画像の内容特定を行うために Web ページ構造を解析することによって近接テキストがどの程度がどこまでの範囲になるのかを検証する手法を提案した。今後の課題としては、より多くの Web ページに対して、検証を行い、本手法の精度向上のためによりタグや近接の定義についての再考察などが挙げられる。

参考文献

[1] 柳井啓司:”キーワードと画像特徴を利用した WWW からの画像収集システム”, 情報処理学会論文誌 Vol.42 No.SIG10(TOD 11),pp79-91,2001

[2] 嶋田和孝, 伊藤哲郎, 遠藤勉:”表層的な情報を用いた画像の内容特定と分類”, 情報処理学会研究報告 2000-NL-139-8,pp55-60,2000

[3] Maro La Cascia, Saratendu Sethi, and Stan Sclaroff:” Combining Texual and Visual Cues for Content-based Image Retrieval on the World Wide Web”, IEEE Workshop on Content - Based Access of Image and Video Libraries June 1998,pp24-28(1998)