ブログ記事を対象とした色情報を用いた画像ラベリング

牛久保 佑樹[†1] 藤田 茂[†2]

インターネットの普及によりウェブ上で多くの画像が見られるようになった．しかし，ウェブ上にある多くの画像はラベリングがなされておらず，画像の検索や整理が困難といった問題点がある．本研究では，形態素解析器を用いたキーワード抽出と，ヒストグラムインターセクションを用いたラベリング手法を提案する．形態素解析器を用いて画像近辺のテキストからキーワードを抽出し，抽出したキーワードを用いて画像検索を行う．次にヒストグラムインターセクションを用いて検索結果の画像とラベリング対象となる画像の比較を行なう．この時，画像の類似度が高ければ，画像検索の際に使用したキーワードはラベリング対象となる画像との関連性が高いと判断できる．この操作を抽出した全てのキーワードに対して行ない，最も類似度の高いキーワードで画像のラベリングを行う．本手法を用いて 100 件のブログ記事を対象に実験を行った結果，ラベリング精度は 22%となった．

Color histgram based image labeling for weblogs

Yuki Ushikubo[†1] and Shigeru Fujita[†2]

In this paper, We propose a method with extraction keyword by morphological analyzer and color histgram based comparison for image labeling. We extract keyword from text near the image, and we get a image using by extracted keyword from image search results. Next, we compare labeling target image and image search results on the internet. If similarity of images is high score, evtracted keyword and labeling target image has high relevance. We do this method for all extracted keyword and all labeling target image, and we name images by most relevant keywords. We have conducted experiments with 100 blog posts, we got 22% accuracy.

†1 千葉工業大学大学院情報科学研究科
Graduate school of Information and Computer Science, Chiba Institute of Technology
†2 千葉工業大学情報科学部
Faclty of Computer and Information Science, Chiba Institute of Technology

1. はじめに

デジタルカメラやイラストレーションソフト，個人ブログの普及により，ウェブ上で多くの画像が見られるようになった．ブログへ画像ファイルをアップロードした際，ファイル名は UUID 等を用いた記号列にされることが多く，ファイル名が画像の内容を表すキーワードであることは少ない．そのため、閲覧者が画像ファイルにラベル付けを行いたい場合には，閲覧者自身の知識からラベル付けを行なうか，ブログ記事から関連キーワードを探し出す必要がある．しかし，人手でブログ記事から関連キーワードを探し出すことは手間がかかるほか，閲覧者にとって未知の画像・キーワードである場合に，関連性を判断することが難しいといった問題点がある．この問題を解決する手法として，HTML タグの表す意味を元に画像に関連するキーワードを抽出する手法[1] や，HTML テキストの重要文から画像の内容を表すキーワードを抽出する手法[2] が提案されている．これらの手法では，HTML タグ中に画像に関するメタデータが含まれている場合や，HTML テキストの重要文中に画像を表すキーワードが含まれている場合に，関連キーワード抽出することができる．しかし，ブログは各ユーザの個人的な嗜好で書かれるため，HTML タグや重要文の中に関連キーワードが含まれていない場合がある．そのため，HTML テキストのみに重点を置いた手法では，関連キーワードを取得できない可能性がある．そこで本研究では，ブログ記事に掲載された画像に関連するキーワードを，色情報を用いた画像比較を行なうことによって，同じブログ記事内のテキストから抽出する手法を提案する．

2. 提案手法

本研究では，形態素解析器を利用したキーワード抽出と，ヒストグラムインターセクションによる画像比較を行うことによって，画像とキーワードの関連性を測る手法を提案する．提案手法の処理の流れを図 1 に示す．

提案手法では，まず形態素解析器 SEN を用いてキーワードの抽出を行う．画像は一般的に，実体のある物や風景を撮影・描画したものがほとんどである．よって，サ変接続名詞など動作を表す名詞はラベリングに使用するキーワードとしては不適当であると考えられる．そこで，今回は一般名詞と固有名詞に限定してキーワードの抽出を行う．

次に，抽出したキーワードで画像検索を行ない，検索結果の画像を画像 DB に保存する．検索には Yahoo!画像検索を使用し，検索結果先頭 20 件の画像を比較用の画像として取得した．

図 1　処理の流れ

```
1  begin
2  blog = getWeblog(); //ブログ記事を取得する
3  tIMG = getImg(blog); //ラベリング対象の画像を取得する
4
5  //blog から取得した keyword が null になるまで繰り返し
6  while((Keyword=extractKeyword(blog))!=null){
7      //keyword を用いた画像検索結果から 20 件の画像を取得
8      sImg[20] = getImgSearchResults20(Keyword);
9      //tImg，sImg をそれぞれ比較しヒストグラムインターセクションの平均値を求める
10     HistValue = HistgramintersectionAverage(tImg,sImg);
11
12     //HistValue 値の最も高い Keyword をラベルとして保存する
13     if(isHiscore(HistValue))
14         label = Keyword;
15 }
16
17 //label で tImg にラベリングを行う
18 setLabelImg(tImg,label);
19 end
```

次に，ヒストグラムインターセクションを用いて，実験対象となる画像と，検索結果の画像の類似度の比較を行なう．このとき，通常のフルカラー画像では色数が約 1、600 万色と多く，同一色の数で類似度判定を行うヒストグラムインターセクションは，そのままでは類似度を測る有効的な手法とはならない．そのため，類似度判定を行う前に画像の色数を 64 色に減色する．ヒストグラムインターセクションの式を以下に示す．なお，R は類似度，Ha は画像 a のヒストグラム値，min(x; y) は両者を比較して小さい方の値を返す関数である．

ふたつの画像が完全一致する場合に $R = 1.0$ の最大値を取り，完全不一致の場合に $R = 0$ となる．

$$R = \frac{\sum_{i=1}^{n} min(H_1[i], H_2[i])}{\sum_{i=1}^{n} (H_1[i])} \tag{1}$$

$R = 1.0$ であれば，実験画像と検索結果の画像は同一のものである．すなわち，画像検索に利用したキーワードは，実験画像と検索結果の画像の両方を表すキーワードである．よって，画像類似度 R が高ければ，抽出したキーワードは画像との関連性が高いと言える．

画像 DB には各キーワードごとに 20 件の画像が格納されている．ヒストグラムインターセクションによる比較を 20 件それぞれの画像に対して行い，その平均値を取る．この操作をラベリング候補となるキーワードすべての画像に対して行い，ヒストグラムインターセクションの値が最も高いキーワードでラベリングを行った．

3. 評 価 実 験

代表的な 5 つのブログサービス 5 つからランダムに各 20 件，合計 100 件の記事を抜きだし提案手法でどの程度の精度が得られるかを確認した．実験結果を表 1 に示す．

表 1　実験結果

対象	画像数（件）	精度（%）
Ameba ブログ	20	14
FC2 ブログ	20	29
Goo ブログ	20	17
Livedoor ブログ	20	26
Yahoo!ブログ	20	23
平均	20	22

4. お わ り に

本研究では，色情報を用いた画像比較を行なうことによって，画像に関連するキーワードを抽出する手法を提案した．今回対称にしたブログでは平均で 22%の抽出率でキーワードを抽出した．

抽出率が変化する要因として，今回の研究では画像比較に全域的な色情報を利用しているため，似たような色味を持つ複数のキーワードを抽出した場合や，カメラ画像の日照変化による変化などで，抽出率が変動するものと考えられる．

参 考 文 献

1) E.V.Muson and Y.Tsymbalenko: To search for images on the web, lood at the text, then look at the images, *First Internatinal Workshop on Web Document Analysis*, pp.39–41 (2001).
2) 相良直樹，砂山　渡，谷内田正彦：HTML テキストの重要文を用いた画像ラベリング手法，電子情報通信学会論文誌，No.2, pp.145–153 (2004).