1F1-07

# Web 上の画像情報源発見に向けて

Towards the Web Image Resource Mining

重森圭輔[1] Keisuke Shigemori
Zoran Stejic[2] Zoran Stejic
廣田薫[2] Kaoru Hirota
山口亨[1] Toru Yamaguchi
高間康史[1] Yasufumi Takama

[1]東京都立科学技術大学 Tokyo Metropolitan Institute of Technology
[2]東京工業大学 Tokyo Institute of Technology

Abstract: As the web is vast and disorderly, it is difficult to find desired information on the web. In particular, finding image resources (knowing where and what kind of images can be found on the web) is very difficult but challenging. As the first step towards the web resource mining, this paper reports the preliminary results of collecting image information by a web robot as well as presenting meta information of images.

## 1. はじめに

近年，ハードウェアテクノロジーの発展により低コストでの大容量データの蓄積も可能になってきた．しかし，その結果Web上での情報の無秩序さはますます増大してきている．しかし，我々はその無秩序さの中から有益な情報を集めてこなければならない．本研究ではWeb上の情報，特に画像情報源の発見を目的とする．

## 2. Web 画像情報源の収集

Web上の画像およびサイトの特性を見るために以下のような手法を用いた．

(1) Webロボットにより Webページ，画像および画像に関するメタ情報を収集
(2) 収集した画像に関するメタ情報の提示
→サイトごとの画像情報の傾向を見る．
(3) LSPを利用したサイト集約画像の提示
→サイトごとの画像の視覚的特徴を見る．
(4) サイト間リンク状況の可視化
→画像サイトの繋がりを視覚的に提示．

## 3. Web ロボットによる画像メタ情報の収集

本研究では必要とするデータをまずWeb上からダウンロードしてこなければならない．そこで画像を自動収集してくるWebロボットを作成した．

Webロボットはインターネット上にあるページのリンクを辿り，情報を収集してくるプログラムである[1]．今回はリンクを辿り，ページと画像,そして画像に関するメタ情報を収集するプログラムを作成した．

探索の方法は幅優先探索と深さ優先探索が一般的であるが，今回はサイトごとに画像を集めやすく，区切りもつけやすいという理由から深さ優先探索を選択し，５階層まで収集を行った．

SEEDページ(Yahoo!のカテゴリー)を与え,画像は出現頻度とLSP[4]に利用することを考えJPEGとGIFのみ収集した．収集したデータの概要は表１に示す．

連絡先：東京都立科学技術大学高間研究室，〒191-0065 東京都日野市旭が丘六丁目６番地，Tel: 042-585-8629，shigemori@krectmt3.tmit.ac.jp

収集した画像に関するメタ情報は以下の通りである

- 画像本体
- 画像の保存名称
- Web上での画像の名称
- 収集日時
- 画像が使用されていたページのURL
- 画像が実際に置いてあったURL

表１：収集したデータの概要

| | | Data1 | Data2 |
|---|---|---|---|
| 内容 | | 写真家 | イラストレーター |
| ページ数 | | 5450ページ | 2947ページ |
| サイト数 | | 228サイト | 124サイト |
| 画像総数 | | 23811枚 | 12445枚 |
| 画像総サイズ | | 366,189,855 | 111,571,679 |
| サイト | 最大画像枚数 | 2185枚 | 2286枚 |
| | 平均画像枚数 | 104.4枚 | 100.4枚 |
| | 平均画像容量 | 1,606,096 | 889,772 |
| 単体 | 平均サイズ | 15,379 | 8,965 |
| | JPEG平均 | 27,386 | 22,137 |
| | GIF平均 | 4,676 | 6,395 |
| 割合 | 10kByte以上の画像数 | 7579枚 (31.8%) | 2728枚 (21.9%) |
| | 直リンク画数 | 437枚(1.8%) | 29枚(0.2%) |
| | JPEGとGIFの割合(枚数) | J 11222枚 (47.1%) G 12589枚 (52.9%) | J 2032枚 (16.3%) G 10413枚 (83.7%) |
| | JPEGとGIFの割合(サイズ) | J 307,324,225 (83.9%) G 58,865,630 (16.1%) | J 44,982,352 (40.3%) G 66,589,327 (59.7%) |

※サイズの単位はByteである．

SEEDページ

Data1:(http://dir.yahoo.co.jp/Arts/Visual_Arts/Photography/Photographers/)
Data2:(http://dir.yahoo.co.jp/Arts/Visual_Arts/Illustration/Illustrators/)

## 4. サイトごとの画像の特徴

画像検索において，現在利用されているキーワードを使った画像検索だけではなく，画像自体を利用した検索方法が必要となってきている．近年では近似画像を利用した画像検索が数多く研究されているが、検索効率の向上が課題となっている[2][3]．

これらの問題はクラスタリングなどを用いて検索対象を狭める事で改善することができるが，さらにどの様な画像が対象データベースに含まれるのかを可視化し，検索者に提示することも有効であると考える．今回はその第一歩として，LSP を利用してサイトごとにその内容を集約する画像を作成することを試みる．

### LSP (Local Similarity Pattern)

LSP は Stejic らの提案する，検索者の観点を考慮可能な画像類似性判別手法である．複数の正例画像から，図１のような共通する特徴量を GA を用いて抽出する．ここで，C,S,T と記されたブロックではそれぞれ色，形状，テクスチャがその領域の特徴量としてふさわしいことを示している．

| | | | |
|---|---|---|---|
| C | T | C | T |
| S | C | C | T |
| ST | C | CS | S |
| T | S | ST | ST |

C:COLOR（色）
S:SHAPE（形状）
T:TEXTURE（模様）

図１：LSP によって抽出された特徴量

LSP を使うにはサイトから正例として用いる画像を選ばなければならない．今回は以下のような考え方から２種類の方法で正例画像 20 枚を選択した．

（１）大きい画像の方がサイトを代表している．
　　→サイズの大きい順に選択
（２）同じサイトならどの画像も同じである．
　　→ランダムに選択

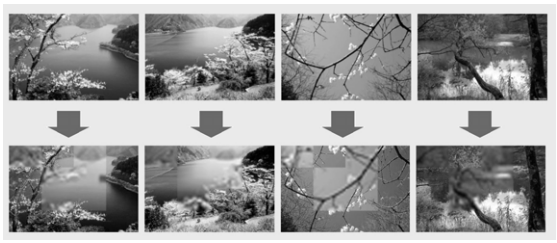

図２：LSP を使ったサイト集約画像(1)

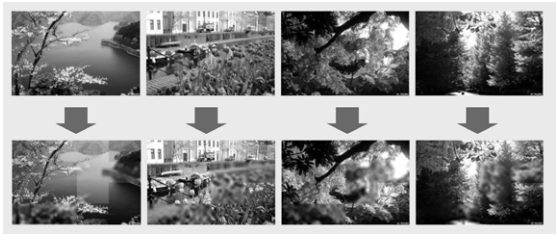

図３：LSP を使ったサイト集約画像(2)

図２，３は作成した LSP に基づき，C 以外のブロックは白黒に，S,T 以外のブロックは解像度を落とした画像である．

図２はサイズの大きい順に，図３はランダムに選んだ代表画像から作ったサイト集約画像である．サイズ順で選んだもののほうが，正例画像として似た画像が選び出されるため，意味のある LSP を作り出せるようである．

## 5. サイト間のリンク関係

Web 上の画像情報源を発見するためには，どのサイトにどれくらいの画像があり，どのように繋がっているかを調べる必要があると考え，バネモデルに基づく情報可視化インタフェースを用いてユーザに提示する．

バネ長はサイト A から B へのリンクを a，B から A へを b としたとき log(a+b+1)で計算し，その後，正規化している．バネモデルの可視化には我々の開発した情報可視化システム TMIT を利用した[5]．オブジェクトは先頭にある数字がそのサイトの画像の枚数，続いてサイトのドメイン名である．

Data2 を可視化すると図４のようになる．図４の中央にあるサイト「www.idea.gr.jp」はイラストライブラリであると同時に画像サーチエンジンも内在するサイトであるため，他の多くのサイトとリンクが多く，中央に配置されている．

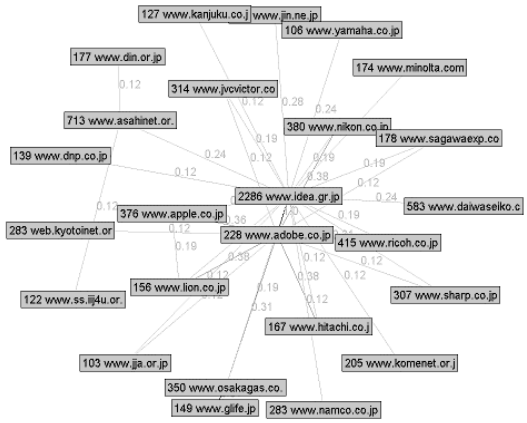


図４：リンク数を元にしたサイト間の関係

## 6. おわりに

本研究でサイトごとの画像の特徴とサイトごとのリンク関係をユーザに提示することができた．今後の課題としては，LSP に重さを加えたサイト集約画像の作成，サイト単位による画像検索の改善，web からの画像データベースの構築などが挙げられる．

## 7. 参考文献

[1] 山田誠二，村田剛志，北村靖彦，知的 Web 情報システム，人工知能学会誌，Vol.1，No.4，pp.495-502，2001.

[2] 鷲澤輝芳，矢田徹，安田靖彦，画像データベース類似検索における k 最近傍探索の高速計算アルゴリズム，NII Journal，NO.2，pp.27-37，2001.

[3] 北本朝展，高木幹雄，類似画像検索への応用を目的とした階層化属性付きグラフマッチングの高速化，画像の認識理解シンポジウム(MIRU’96)，Vol.2，pp.331-336，1996.

[4] Zoran Stejic, Yasufumi Takama, Kaoru Hirota, Genetic algorithm-based relevance feedback for image retrieval using local similarity patterns，Information Processing & Mnagement，Vol.39，pp.1-23，2003.

[5] Yasufumi Takama, Tetsuya Hori, Application of Immune Network Metaphor to Keyword Map-based Topic Stream Visualization, CIRA2003, (to appear)